FACULTY OF PSYCHOLOGY
AND EDUCATIONAL SCIENCES

Reduction in sample size by order restrictions

Leonard Vanbrabant

Supervisor: Prof. Dr. Yves Rosseel
Co-supervisor: Dr. Rens van de Schoot

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Doctor of Psychology

Academic year 2017–2018

GHENT
UNIVERSITY

To my wife Juul and my sons Tijl and Peer

Table of Contents

Expression of gratitude

First, I would like to thank and express my sincere gratitude to my supervisor Yves Rosseel and co-supervisor Rens van de Schoot for their continuous support of my PhD. Yves, fortunately, you gave me all the opportunities to explore my own research interests. In doing so, I could investigate research areas in statistics I didn't know even existed. This exploration has really broaden my view and way of thinking. I have learned a lot from you. Not only about statistics and writing your own software package, but also that good food (especially Indian) and Belgian beer are two very important ingredients when talking about statistics. I really enjoyed those moments. Rens, you always ensured that I kept my sanity during the many sleepless nights as a young father. The many interesting discussions we had while eating a delicious burger helped me going. So, thank you for that. Both of you provided substantial support and feedback for all of my papers. Without your support, this thesis would not have been in its present form.

Thanks also go to my guidance committee members – Prof. dr. Beatrijs Moerkerke, Prof. dr. Stijn Vansteelandt and Prof. dr. Herbert Hoijtink for their feedback and sound advices. A special thanks goes to our secretary Isabelle. You were always very helpful to me.

I also want to thank my friends for their unconditional support. I really enjoyed our talks, playing tennis, drinking beer and eating pizza, organizing CSI meetings and going to conferences. Special thanks go to Jeroen for helping me out with the CSS of my website and to Jesse for helping me converting Fortran code to R code, which was a real pain in the neck.

Last but not least, a special gratitude and love goes to my family for their support. Special thanks go to my lovely wife Juul (sorry for working in the evenings and weekends) and to my two awesome sons Tijl and Peer, who have enriched my life immensely and without whom this dissertation would have been completed a year earlier.

Eindhoven, 14 december 2017 *Leonard Vanbrabant*

1 Introduction

Researchers often have substantive research questions that involve informative hypotheses. Consider, for example the following typical examples:

1. Cognitive behavioral therapy (CBT) in combination with drugs is *more* effective against depression than CBT only; in addition, the new drug is *more* effective than the old drug.

2. Facial burns would have a *higher* impact on self-esteem than body burns and the impact for both types would be *higher* in females than in men.

3. There is a *positive relation* between social skills, interest in artistic activities and use of complicated language patterns, and the target variable IQ.

4. The exercises (no training, physical training, behavioral training, and a combination of physical and behavioral therapy) are associated with a *reduction* in the mean aggression levels.

These hypotheses are called informative because they include directional expectations about the ordering of the parameters. For example, in the

first hypothesis a clear ordering between the three drug treatment means is expected and in the third hypothesis the regression coefficients for social skills, interest in artistic activities and use of complicated language patterns are expected to be positively related to IQ. This prior knowledge originates from previous research (i.e. theory) or academic reasoning and can be translated into an order-constrained hypothesis by means of imposing order constraints (i.e. $\leq$, $\geq$, $=$) on the model parameters. Thus, in statistical symbols these four informative hypotheses might be expressed as the following order-constrained hypotheses:

$$
\begin{aligned}
H_1 &: \mu_{\text{new drug}} \geq \mu_{\text{old drug}} \geq \mu_{\text{no drug}} \\
H_2 &: \mu_{\text{men;body}} \leq \mu_{\text{men;face}} \leq \mu_{\text{females;body}} \leq \mu_{\text{females;face}} \\
H_3 &: \beta_{\text{social}} \geq 0, \beta_{\text{artistic}} \geq 0, \beta_{\text{language}} \geq 0 \\
H_4 &: \mu_{\text{no}} \leq \{\mu_{\text{physical}} = \mu_{\text{behavioral}}\} \leq \mu_{\text{combination}},
\end{aligned}
$$

where μ reflects the population mean for each group and β is a regression coefficient.

1.1 Current practice

Classical null-hypothesis significance testing (NHST) is the most widely used method in the social and behavioral sciences to evaluate a hypothesis. To evaluate hypotheses like H_1, H_2 and H_4, we usually use an ANOVA where the hypothesis is tested that all means are equal (nothing is going on) against the alternative unconstrained hypothesis that something is going on. For example, for the hypothesis H_1 the null-hypothesis equals $H_{01} : \mu_{\text{new drug}} = \mu_{\text{old drug}} = \mu_{\text{no drug}}$ and the alternative hypothesis equals $H_{u1} : \mu_{\text{new drug}}\ ,\ \mu_{\text{old drug}}\ ,\ \mu_{\text{no drug}}$. If the resulting F-test is significant, all we know is that some means are not equal and additional contrast tests are needed to find evidence in favor of the hypothesis of interest.

Another frequently used approach for evaluating a directional hypothesis like the ones above is linear trend analysis. To tests whether the three group means in H_1 follow a decreasing order, predefined weights are specified on the means. In case of three groups, the weights +1, 0, and -1 are often used. The contrast compares the lowest group mean with the

highest group mean. Again, if the F-test is significant, all we know is that the linear trend is not zero and additional diagnostics are needed to support the conclusion of a linear trend.

1.2 History, critiques and alternatives

1.2.1 History

NHST as we know it today began with Karl Pearson (Pearson, 1900) who introduced the chi-squared test of goodness of fit, and the p-value associated with this test-statistic. This was followed by Willam Gosset's (pseudonym: Student) discovery of the t-distribution (Student, 1908). However, it was Fisher (Fisher, 1925) who popularized significance tests and p-values. The theory of Fisher was further 'improved' by Neyman and Egon Pearson (Neyman & Pearson, 1928) who introduced hypothesis testing.

Fisher's approach is to use the data to provide evidence for the null-hypothesis. No alternative hypothesis exists and it is the null-hypothesis that is to be nullified. Note that the null-hypothesis does not need to be a zero difference. In Fisher's approach (Gigerenzer, 2004) the researcher sets up a null-hypothesis that a sample comes from a population with a known sampling distribution (e.g. t-distribution). The null-hypothesis is disproved if the sample estimate is as extreme or more extreme than we would expect by chance. Fisher regarded the p-value as inductive evidence against the null-hypotheses. The smaller the p-value, the more convincing the evidence against the null-hypothesis. The researcher is supposed to decide if the evidence is convincing enough but does not talk about accepting or rejecting the hypothesis.

Some authors have argued that the theory of Fisher is defective because the null-hypothesis cannot be rejected without providing evidence for another (i.e. alternative) hypothesis (Sober, 2008). Specification of an alternative hypothesis is the key difference between Fisher's and Neyman-Pearson's methodologies. Although Fisher used some kind of alternative when computing a p-value, he never explicitly defined nor used specific alternative hypotheses. With the specification of an alternative hypothesis, Neyman and Pearson added concepts of Type-II error rates (β), and relatedly, statistical power. In Neyman-Pearson's approach (Gigerenzer, 2004)

the researcher sets up a null-hypothesis and an alternative hypothesis, and decides about α, β, and the sample-size (power calculations) a priori to the experiment. These define the rejection region. Then, if the data falls into the rejection region of the null-hypothesis, the null-hypothesis is rejected. Otherwise the null-hypothesis is accepted. Note that accepting a hypothesis does not mean that you believe in it but only that you act as if it were true.

NHST is considered as a compromise between Fisher's theory on significance testing and the concepts from Neyman-Pearson. However, there is not a single agreement upon the characterization of this hybrid NHST (Little, 2013). Some authors have argued that the hybrid logic is a confusing and inconsistent mixture of the two different decision theories (Gigerenzer, 1993). On the other hand, Lehmann (1993), a former student of Neyman, argued that at a practical level, the two approaches are complementary and that p-values, significance levels and power can be combined into a unified approach (Spanos, 2003). The popularity of NHST is probably due to the textbooks written (largely by non-statisticians) in the 1940s to 1960s to teach students in the social sciences the 'rules of statistics' (Gigerenzer et al., 1989). In addition, the hybrid theory was standardized by editors of major journals. Researchers were therefore more or less forced to use significance tests (Morrison & Henkel, 1970).

1.2.2 Critiques

NHST has survived many attacks since its introduction in the 1940s (see Nickerson, 2000 and the references therein). One of the main critiques is that the hypothesis of interest cannot be tested directly. Reconsider the order-constrained hypothesis H_1. To evaluate this directional hypothesis using an ANOVA, the null-hypothesis H_{01} is tested against the alternative hypothesis H_{u1}. Obviously, the hypothesis of interest H_1 is not part of the null-hypothesis or the alternative hypothesis. Consequently, the resulting F-test does not capture the a-priori ordering of the means and additional contrast tests are required to find evidence in favor of H_1. The aftermath would be an inflated Type-I error rate (α), or a decrease in power when an α correction is used. In addition, using linear contrast tests to test a directional hypothesis may result in spurious conclusions with regard to the direction of the effect. For example, if the sample means for H_1 are

5, 10, and 1, the contrast test with weights +1, 0 and -1 will probably reject the null-hypothesis in favor of a nonzero linear trend, even though the first order constraint is violated.

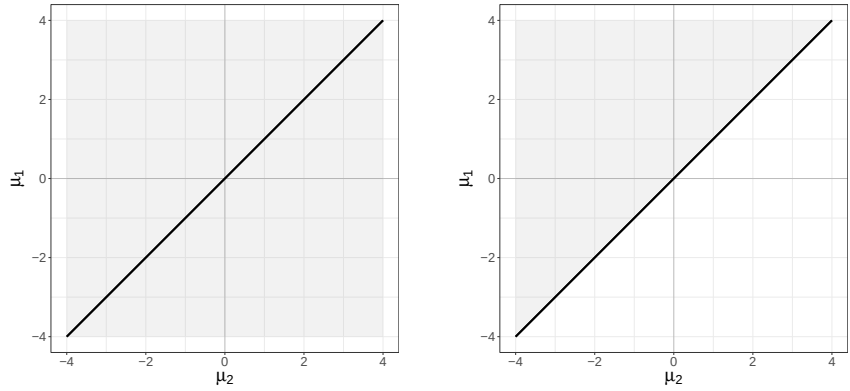
1.2.3 Alternatives

Regardless the numerous critiques raised against NHST for the past 80 years, it is still the most taught decision theory in undergraduate courses. As a result, it has slowed down scientific progress. The best option seems to abandon NHST and to start teaching alternative available methods, among them effect sizes (Cohen, 1988), confidence intervals (Neyman, 1935), meta-analysis (Rosenthal, 1984), Bayesian hypotheses testing (Lindley, 1965) and model selection using information criteria (e.g., Akaike, 1998). In this dissertation, we will investigate yet another alternative, i.e. constrained statistical inference or informative hypothesis testing (e.g., Hoijtink, 2012; Kuiper, 2011; Silvapulle & Sen, 2005. Since this is key to this dissertation, we will further elaborate on this.

1.3 Informative approaches

Since the early 1950s a vast amount of literature has been produced in both the frequentist framework (Barlow, Bartholomew, Bremner, & Brunk, 1972; Kuiper, 2011; Robertson, Wright, & Dykstra, 1988; Silvapulle & Sen, 2005) and in the Bayesian framework (e.g., Hoijtink, 2012) for evaluating informative hypotheses such as $H_1 - H_4$ directly. To evaluate an informative hypothesis, three methods can be distinguished, i.e. hypothesis testing, model selection using information criteria and Bayesian model selection. For an overview and a comparison see Hoijtink and Klugkist (2007), Kuiper and Hoijtink (2010) and Van de Schoot, Hoijtink, and Romeijn (2011).

The advantage of informative hypotheses compared to classical NHST is that the hypothesis of interest can be evaluated more directly. Consequently, substantial smaller samples are needed to detect specific effects. Non-technically, this is because the parameter space is restricted and it is easier to find evidence for or against a smaller parameter space compared to finding evidence for a larger parameter space. To illustrate, consider Figure 1.1a, where the parameter space is defined by $H_{u5} : \mu_1, \mu_2$ (no



(a) $H_{05} : \mu_1 = \mu_2$ vs. $H_{u5} : \mu_1,\ \mu_2$. (b) $H_{05} : \mu_1 = \mu_2$ vs. $H_5 : \mu_1 \geq \mu_2$.

Figure 1.1: The admissible area is shaded gray.

constraints are imposed on the means). Note that we only depicted the parameter space between -4 and 4 and not the whole parameter space. Then, the unrestricted parameter space consists of all possible values for both parameters. In other words it consists of the entire range of admissible hypotheses. Next, consider Figure 1.1b, where the parameter space is now restricted by the order constraint $H_5 : \mu_1 \geq \mu_2$. Now not all possible combinations between μ_1 and μ_2 are admissible. Therefore, the range of possible statistical hypotheses is also smaller because only the combinations in according with H_5 are allowed. Therefore it is easier to differentiate the order-constrained hypothesis from the alternative hypothesis. Hence, a higher power is gained and consequently a smaller sample-size is needed. As we will show in the next Chapter, this sample-size reduction may be as high as 50%.

Informative methods have demonstrated real value in improving NHST, but unfortunately these methods are rarely used in the social and behavioral sciences. The absence of these methods in a researcher's toolbox can be understood on three levels, i.e. textbook writers, benefits of alternative methods, and software. On the first level, writers of todays textbooks for the social and behavioral sciences hardly mention alternatives for NHST. For instance, checking three books on statistics for the behavioral and

social sciences (Gravetter & Wallnau, 2014; Sirkin, 2005; Tabachnick & Fidell, 2007) that were readily available, I found that they barely or not at all hinted on the controversy of NHST. On the second level, today's researchers are trained at a time that NHST is the predominant method of statistical inference. Hence, researchers are probably unaware of the (major) benefits of informative hypotheses. On the third level, no (user-friendly) software tool exists that can deal with order constraints in a variety of statistical models. The available software tools are scattered and limited to ordered means and variances, and to ordered regression coefficients in a linear model.

1.4 Objectives

In this dissertation, we cover three main topics, i.e. reduction in sample-size, model selection using order-constrained information criteria, and software. The first objective is to investigate the reduction in sample-size (gain in power) when an increasing number of order constraints is imposed on the means of an ANOVA and on the regression coefficients of a linear model. In addition, we also investigate the effects of outliers on the power. The second objective is to introduce an alternative method to order-constrained hypothesis testing for evaluating an order-constrained hypothesis against its complement using information criteria. The third objective is to develop software tools for estimating and evaluating order-constrained hypotheses for a variety of statistical models. In addition, we provide a clear tutorial on how these tools can be used to evaluate informative hypotheses. Next, we will discuss each of these topics in more detail.

1.4.1 Sample-size

There are three basic testing problems that can be considered in connection with informative hypotheses. In the literature they are often called hypothesis test Type A, hypothesis test Type B (Silvapulle & Sen, 2005). and hypothesis test as Type C. In this dissertation, we shall consider solely hypothesis test Type A and hypothesis test Type B. The role of hypothesis test Type C is merely to complete the set of tests. Its practical use is limited because its power is quite low (Grömping, 2010). In words, these

hypothesis tests can be defined as follows:

Type A test: H_{A0} : all restrictions are active (=)
vs. H_{A1} : at least one order restriction is strictly true (>)

Type B test: H_{B0} : all restrictions hold in the population
vs. H_{B1} : at least one restriction is violated

Type C test: H_{C0} : at least one restriction is false or active (=)
vs. H_{C1} : all restrictions are strictly true (>)

In the null-hypothesis H_{A0} of hypothesis test Type A all order constraints are treated as equality constraints and is tested against the alternative order-constrained hypothesis H_{A1}. In hypothesis test Type B, the null-hypothesis H_{B0} is the order-constrained hypothesis and is tested against its complement H_{B1}. In hypothesis test Type C, the alternative hypothesis consists of strict order constraints only and is tested against the null-hypothesis that at least one order constraint is violated.

To find evidence in favor of an order-constrained hypothesis, we use a combination of hypothesis test Type B and hypothesis test Type A (in this order), which we call hypothesis test Type J. The rationale is that if hypothesis test Type B is *not* significant, we do not reject the null-hypothesis that all restrictions hold in the population. However, hypothesis test Type B cannot make a distinction between inequality and equality constraints. Therefore, if hypothesis test Type B is not significant, the next step is to evaluate hypothesis test Type A. If we reject its null-hypothesis H_{A0}, we can conclude that at least one inequality constraint is strictly true. Then, if we combine the evidence of hypothesis test Type B and hypothesis Type A, we can say that we have found indirect evidence in favor of (or against) the order-constrained hypothesis. A measure of effect-size can aid in the interpretation of the strength of this support.

In Chapter 2, we study the relationship between order constraints and sample-size for hypothesis test Type J. More precise, by means of a simulation study we investigate the reduction in sample-size when an increasing number of order constraints is imposed on the means of an ANOVA and on the regression coefficients of a linear regression model. The main re-

sults are power tables for hypothesis test Type J. These power tables are comparable with the familiar power tables in (Cohen, 1988) which are seen as the 'gold' standard. The major advantage of our power tables is that researchers can look up the necessary sample-size with predefined power of 0.80 *and* predefined number of order constraints. In addition, we developed a software tool because the power tables for order-constrained tests only cover a subset of all possible models, while the software tool can be used for all possible combinations.

1.4.2 Outliers

In inferential data analysis, a problem that is often ignored is that data collected may contain irregularities that deviate from the majority of the data, such as outliers in the response space. Consider, for example the simple linear regression example in Figure 1.2, where the data contains one outlier in the response space. The ordinary least squares (OLS) estimator, which is usually used in ANOVA and linear regression is in this case unduly influenced by a single outlier (see solid black line). This results in biased estimates and a decline in statistical power. Fortunately, to deal with these issues, various robust estimators haven been proposed, such as the commonly used M-estimators (Huber, 1973) and MM-estimators (Yohai, 1987). Again, consider Figure 1.2. Clearly, the response outlier has no impact on the robust estimated regression line (see dashed line).

In Chapter 3, we explore the impact of order-constrained robust and non-robust estimators on the power when the data are contaminated with 10% outliers in both the response variable and predictor variables. This is done by means of a simulation study, where we compare the performance of the order-constrained (non-robust) OLS estimators, and (robust) M-estimators and MM-estimators. An empirical example about child and parental adjustment following a pediatric burn event illustrates the application of these robust tests.

1.4.3 Model selection

Besides hypothesis testing, another method for evaluating order-constrained hypotheses is model selection using information criteria. The advantage of model selection compared to hypothesis testing is that model selection

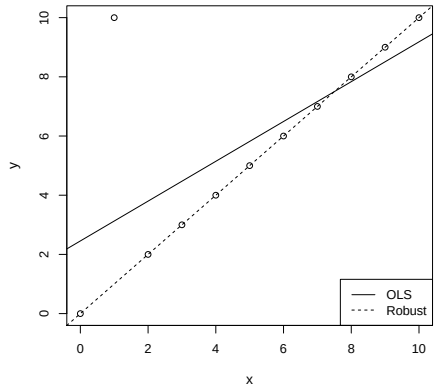


Figure 1.2: The effect of one response outlier on the OLS- and robust-estimator.

has the ability to quantify evidence for the hypothesis of interest. This can be done by computing a relative evidence based on two model probabilities. This relative evidence is simply interpreted as the strength of evidence in favor of one hypothesis over the other (Burnham & Anderson, 2002). The AIC (Akaike, 1998) is probably the most familiar and widely used information criterion employed in the social and behavioral sciences. Nevertheless, the AIC is not suitable when the model parameters are subject to order constraints. A modification of the AIC that can deal with most linear order constraints in multivariate normal linear models is the generalized order-restricted information criterion (GORIC) (Kuiper, Hoijtink, & Silvapulle, 2011).

In Chapter 4, we introduce a method for evaluating an order-constrained hypothesis against its complement H_c using the GORIC (weights). To clarify, reconsider Figure 1.1b, where $H_5 : \mu_1 \geq \mu_2$. Its complement is defined as H_c = not H_5, which corresponds to $H_c : \mu_1 \leq \mu_2$. For the order-constrained hypothesis H_2, the complement is defined as H_c = not H_2. In total, there are 24 ways (i.e., $4! = 4 \times 3 \times 2 \times 1$) in which the four means can be ordered. Hypothesis H_2 consists of 1 of these 24 combinations, therefore the complement represents the 24 - 1 = 23

remaining ways in which the four means can be ordered. An empirical example about facial burn injury illustrates our method.

1.4.4 Software

Although, constrained statistical inference has been around for more than 70 years, software routines are scarce. The available methods are limited, complex, computationally demanding and a user-friendly software routine is often lacking. To fill this gap, in Chapter 5 we present the `R` function `InformativeTesting()` for testing order-constrained hypotheses in structural equation models, which is currently available in the `R` package **lavaan** (Rosseel, 2012). The method uses a likelihood ratio test and the corresponding p-value can be computed based on the parametric bootstrap or Bollen-Stine bootstrap. Since, the p-value can be biased, a double bootstrap procedure is available. Nevertheless, bootstrapping is a computationally demanding procedure, even with today's computer power. Fortunately, in linear regression models this bootstrap procedure can be avoided. Therefore, we developed the `R` package **restriktor** for estimating and evaluating order-constrained hypotheses for regression models. This includes, ANOVA, linear regression, generalized linear regression, robust estimation of the linear regression model and multivariate linear regression. In Chapter 6, we provide a tutorial introduction to restriktor. By means of seven examples we demonstrate how informative hypotheses can be evaluated using both hypothesis tests and model selection using information criteria. More information about restriktor can be found online at `www.restriktor.org`.

It is important to stress that developing an `R` package is not a short-term job. Over the last few years, restriktor has matured from a wobbly single function to a stable comprehensive toolbox for estimating and evaluating order-constrained hypotheses. This means that initial functions have been deprecated over the years and replaced by the restriktor package. To ensure reproducibility of the simulation results and applicability of the examples given in this dissertation, we adapted all `R`-input and output to match the current restriktor version (0.1-70).

1.5 Outline

Chapter 2 In the first study we investigate the gain in power when an increasing number of order constraints is imposed on the means of an ANOVA and on the regression coefficients of a linear model. *The chapter is published as Vanbrabant, L., Van de Schoot, R., & Rosseel, Y. (2015). Constrained statistical inference: sample-size tables for ANOVA and regression. Frontiers in Psychology, 5: 1565.* `http://dx.doi.org/10.3389/fpsyg.2014.01565`.

Chapter 3 In the second study, we compare order-constrained robust and non-robust estimation methods for informative hypotheses. More specifically, we investigate the performance of robust and non-robust estimators in terms of the mean squared error and we investigate the size and power of one-sided robust and non-robust tests.

Chapter 4 In the third study, we introduce a new method on how to evaluate an order-constrained hypothesis against its complement using the GORIC (weights). *This chapter is under revision at Psychological Methods as Vanbrabant, L., Van Loey, N., & Kuiper, R. Giving the complement a compliment: Evaluating an order-constrained hypothesis against its complement using the GORIC.*

Chapter 5 In the fourth study, we present a general method for testing order-constrained hypothesis in structural equation models. *This chapter is published as Vanbrabant, L., Van de Schoot, R., Van Loey, N., & Rosseel, Y. (2017). A General Procedure for Testing Inequality Constrained Hypotheses in SEM. Methodology, 13: 61–70.* `http://dx.doi.org/10.1027/1614-2241/a000123`.

Chapter 6 In this chapter, we demonstrate by seven examples how order-constrained hypotheses can be evaluated using `restriktor`.

Chapter 7 In this chapter, I give thought to my research papers. I discuss the limitations of this dissertation, what could be improved and which topics remain for future research.

Chapter 8 In this chapter, we provide a summary of this dissertation in Dutch.

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), (pp. 199–213). Springer New York: NY. doi: doi:10.1007/978-1-4612-1694-0_15

Barlow, R., Bartholomew, D., Bremner, H., & Brunk, H. (1972). *Statistical inference under order restrictions*. New York: Wiley.

Burnham, K., & Anderson, D. (2002). *Model selection and multimodel inference: a practical information-theoretic approach* (2nd ed.). Spring-Verlag: New York.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Fisher, R. (1925). *Statistical methods for research workers* (6th Ed., 1936 ed.; F. Crew & D. Cutler, Eds.). Oliver and Boyd: Edinburgh, UK.

Gigerenzer, G. (1993). *The superego, the ego, and the id in statistical reasoning* (G. Keren & C. Lewis, Eds.). Hillsdale, NJ: Erlbaum.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. doi: doi:10.1016/j.socec.2004.09.033

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1989). *The empire of chance: How probability changed science and everyday life*. Cambridge University Press. doi: doi:10.1017/CBO9780511720482

Gravetter, F., & Wallnau, L. (2014). *Statistics for the behavioral sciences* (8th ed.). Wadsworth: Belmont CA.

Grömping, U. (2010). Inference with linear equality and inequality constraints using R: The package ic.infer. *Journal of statistical software*, *33*, 1–31. doi: doi:10.18637/jss.v033.i10

Hoijtink, H. (2012). *Informative Hypotheses: Theory and Practice for Behavioral and Social Scientists*. Boca Raton, FL: Taylor & Francis.

Hoijtink, H., & Klugkist, I. (2007). Comparison of hypothesis testing and bayesian model selection. *Quality & Quantity*, *41*, 73–91.

Huber, P. (1973). Robust regression: asymptotics, conjectures and monte carlo. *The annals of statistics*, *1*(5), 799–821. doi: doi:doi:10.1214/aos/1176342503

Kuiper, R. (2011). *Model selection criteria: How to evaluate order restrictions* (Dissertation, Utrecht University). Retrieved from `https://dspace.library.uu.nl/handle/1874/224499`

Kuiper, R., & Hoijtink, H. (2010). Comparisons of means using exploratory and confirmatory approaches. *Psychological Methods*,

15(1), 69–86. doi: doi:10.1037/a0018720

Kuiper, R., Hoijtink, H., & Silvapulle, M. (2011). An akaike-type information criterion for model selection under inequality constraints. *Biometrika*, *98*(2), 495–501. doi: doi:10.1093/biomet/asr002

Lehmann, E. (1993). The fisher, neyman-pearson theories of testing hypotheses: One theory of two? *Journal of the American Statistical Association*, *88*(424), 1242–1249. doi: doi:10.2307/2291263

Lindley, D. (1965). *Introduction to probability and statistics from a bayesian viewpoint, part 1: Probability; part 2: Inference.* Cambridge University press: Cambridge.

Little, T. (2013). *The oxford handbook of quantitative methods: Foundations* (P. Nathan, Ed.). Oxford University Press, Inc.: New York.

Morrison, D. E., & Henkel, R. E. (1970). *The significance test controversy.* Aldine: Oxford, England.

Neyman, J. (1935). On the problem of confidence intervals. *The Annals of Mathematical Statistics*, *6*(3), 111-116.

Neyman, J., & Pearson, E. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: part i. *Biometrika*, *20A*(1/2), 175–240. doi: doi:10.2307/2331945

Nickerson, R. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301. doi: doi:10.1037/1082-989X.5.2.241

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *50*(302), 157–175. doi: doi:10.1080/14786440009463897

Robertson, T., Wright, F. T., & Dykstra, R. L. (1988). *Order Restricted Statistical Inference.* New York: Wiley.

Rosenthal, R. (1984). *Meta-analytic pprocedure for social research.* Sage Publications, Newbury Park.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. Retrieved from `http://www.jstatsoft.org/v48/i02/`

Silvapulle, M., & Sen, P. (2005). *Constrained statistical inference: Order, inequality, and shape restrictions.* Hoboken, NJ: Wiley.

Sirkin, R. (2005). *Statistics for the social sciences* (3th ed.). Sage Publications.

Sober, E. (2008). *Evidence and evolution. the logic behind the science.*

Cambridge University Press.

Spanos, A. (2003). *Probability theory and statistical inference: Econometric modeling with observational data*. Cambridge University Press.

Student. (1908). The probable error of a mean. *Biometrika*, *6*(1), 1–25. doi: doi:10.2307/2331554

Tabachnick, B., & Fidell, L. (2007). *Using multivariate statistics* (5th ed.). Allyn and Bacon: New York.

Van de Schoot, R., Hoijtink, H., & Romeijn, J. (2011). Moving beyond traditional null hypothesis testing: evaluating expectations directly. *Frontiers in Psychology*, 1–5. doi: doi:10.3389/fpsyg.2011.00024

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

2

Constrained statistical inference: sample-size tables for ANOVA and regression[1]

Researchers in the social and behavioral sciences often have clear expectations about the order/direction of the parameters in their statistical model. For example, a researcher might expect that regression coefficient β_1 is larger than β_2 and β_3. The corresponding hypothesis is H: $\beta_1 > \{\beta_2, \beta_3\}$ and this is known as an (order) constrained hypothesis. A major advantage of testing such a hypothesis is that power can be gained and inherently a smaller sample size is needed. This article discusses this gain in sample size reduction, when an increasing number of constraints is included into the hypothesis. The main goal is to present sample-size tables for constrained hypotheses. A sample-size table contains the necessary sample-size at a prespecified power (say, 0.80) for an

[1]This chapter is published as Vanbrabant, L., Van de Schoot, R., & Rosseel, Y. (2015). Constrained statistical inference: sample-size tables for ANOVA and regression. Frontiers in Psychology, 5: 1565. http://dx.doi.org/10.3389/fpsyg.2014.01565.

increasing number of constraints. To obtain sample-size tables, two Monte Carlo simulations were performed, one for ANOVA and one for multiple regression. Three results are salient. First, in an ANOVA the needed sample-size decreases with 30% to 50% when complete ordering of the parameters is taken into account. Second, small deviations from the imposed order have only a minor impact on the power. Third, at the maximum number of constraints, the linear regression results are comparable with the ANOVA results. However, in the case of fewer constraints, ordering the parameters (e.g., $\beta_1 > \beta_2$) results in a higher power than assigning a positive or a negative sign to the parameters (e.g., $\beta_1 > 0$).

2.1 Introduction

Suppose that a group of researchers is interested in the effects of a new drug in combination with cognitive behavioral therapy (CBT) to diminish depression. One of their hypothesis is that CBT in combination with drugs is more effective than CBT only and that the new drug is more effective than the old drug. In symbols this hypothesis can be expressed as H_{CBT}: $\mu_1 < \mu_2 < \mu_3$ ($\mu_1 = CBT_{new\_drug}$, $\mu_2 = CBT_{old\_drug}$, $\mu_3 = CBT_{no\_drug}$), where μ reflects the population mean for each group. To replace the old drug with the new one, the researchers want at least a medium effect size of $f = 0.25$. Classical sample-size tables based on the F test (see for example Cohen, 1988) show that in case of three groups, $f = 0.25$ and a significance level of $\alpha = 0.05$, 159 subjects are necessary to obtain a power of 0.80. However, the expected ordering of the means is in this case completely ignored. When the order is taken into account (here two order constraints), then the results from our simulation study (see Table 2.1, to be explained below) show that with fully ordered means a sample-size reduction of about 30% can be gained.

Consider another example of a constrained hypothesis but now in the context of linear regression. Suppose that a group of researchers wants to investigate the relation between the target variable IQ and five exploratory variables. Three exploratory variables are expected to be positively associated with an increase of IQ, while two are expected to be negatively associated:

- social skills ($\beta_1 > 0$)

- interest in artistic activities ($\beta_2 > 0$)
- use of complicated language patterns ($\beta_3 > 0$)
- start walking age ($\beta_4 < 0$)
- start talking age ($\beta_5 < 0$)

To test this hypothesis an omnibus F test is often used, where the user-specified model (including all predictors) is tested against the null model (including an intercept only). In our example, the null hypothesis is specified as H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$. Classical sample-size tables show that in case of a medium effect-size ($f^2 = 0.10$) 135 subjects are necessary to obtain a power of 0.80 ($\alpha = 0.05$). However, all information about the expected direction of the effects is completely ignored. When this information is taken into account, then our simulation results (see Table 2.2, to be explained below) show that with imposing five inequality constraints, a sample-size reduction of about 34% can be gained. If we impose 2 inequality constraints, the reduction drops to about 14%. This clearly shows that imposing more inequality constraints on the regression coefficients results in more power. Note that the researchers only imposed inequality constraints on the variables of interest. But, this does not have to be the case. Additional power can be gained by also assigning positive or negative associations to control variables. For example, the researchers could have controlled for socioeconomic status (SES). Although, SES is not part of the researchers main interest, they could have constrained SES to be positively associated with IQ if they have clear expectations about the sign of the effect. In this vein, *a priori* knowledge about the sign of a regression parameter can be an easy solution to increase the number of constraints and, therefore, decreasing the necessary sample-size Hoijtink (2012).

Constrained statistical inference (CSI) has a long history in the statistical literature. A famous work is the classical monograph by Barlow, Bartholomew, Bremner, and Brunk (1972), which summarized the development of order constrained statistical inference in the 1950s and 1960s. Robertson, Wright, and Dykstra (1988) captured the developments of CSI in the 1970s to early 1980s and Silvapulle and Sen (2005) present the state-of-the-art with respect to CSI. Although, a significant amount of new developments have taken place for the past 60 years, the relation-

ship between power and CSI has hardly been investigated. An appealing feature of constrained hypothesis testing is that, without any additional assumptions, power can be gained (Barlow, Bartholomew, Bremner, & Brunk, 1972; Bartholomew, 1961a, 1961b; Kuiper & Hoijtink, 2010; Kuiper, Nederhoff, & Klugkist, 2011; Perlman, 1969; Robertson, Wright, & Dykstra, 1988; Silvapulle & Sen, 2005; Van de Schoot & Strohmeier, 2011; Wolak, 1989). Many applied users are familiar with this fact in the context of the classical t-test. Here, it is well-known that the one-sided t-test (e.g., $\mu_1 = \mu_2$ against $\mu_1 > \mu_2$) has more power than the two-sided t-test (e.g., $\mu_1 = \mu_2$ against $\mu_1 \neq \mu_2$), because the p value for the latter case has to be multiplied by two. We show that this gain in power readily extends to the setting where more than one constraint can be imposed. For example, in an ANOVA with three groups the number of order constraints may be one or two, depending on the available information about the order of the means. Hence, we present sample-size tables for constrained hypothesis tests in linear models with an increasing number of constraints. These tables will be comparable with the familiar sample-size tables in Cohen (1988) which are often seen as the 'gold' standard. The major advantage of our sample-size tables is that researchers are able to look up the necessary sample size for various numbers of imposed constraints.

The remainder of this article is organized as follows. First, we introduce hypothesis test Type A and hypothesis test Type B, which are used for testing constrained hypotheses. Second, we present sample-size tables for order-constrained ANOVA, followed by sample-size tables for inequality-constrained linear regression models. For both models we present sample-size tables which depict the necessary sample size at a power of 0.80 for an increasing number of constraints. Next, we provide some guidelines for using the sample-size tables. Finally, we demonstrate the use of the sample-size tables based on the CBT and IQ examples and we provide `R` (R Development Core Team, 2016) code for testing the constrained hypotheses. Note that the article has been organized in such a way that the technical details are presented in the Appendices and can be skipped by less technical inclined readers who are interested primarily in the sample-size tables.

2.2 Hypothesis test Type A and Type B

In the statistical literature, two types of hypothesis tests are described for evaluating constrained hypotheses, namely hypothesis test Type A and Type B (Silvapulle & Sen, 2005). A formal definition of hypothesis test Type A and hypothesis test Type B is given in Appendix A. Consider for example the following (order) constrained hypothesis: H: $\mu_1 < \mu_2 < \mu_3$. Here, the order of the means is restricted by imposing two inequality constraints. In hypothesis test Type A, the classical null hypothesis H_{A0} is tested against the (order) constrained alternative H_{A1} and can be summarized as:

Type A:

$$\begin{aligned} H_{A0}: &\quad \mu_1 = \mu_2 = \mu_3 \\ H_{A1}: &\quad \mu_1 < \mu_2 < \mu_3 \ . \end{aligned} \tag{2.1}$$

In hypothesis test Type B, the null hypothesis is the (order) constrained hypothesis H_{B0} and it is tested against the two-sided unconstrained hypothesis H_{B1} and can be summarized as:

Type B:

$$\begin{aligned} H_{B0}: &\quad \mu_1 < \mu_2 < \mu_3 \\ H_{B1}: &\quad \mu_1 \neq \mu_2 \neq \mu_3 \ . \end{aligned} \tag{2.2}$$

Note the difference with classical null hypothesis testing, where the hypothesis H_{A0} is tested against the two-sided unconstrained hypothesis H_{B1}. To evaluate constrained hypotheses, like H: $\mu_1 < \mu_2 < \mu_3$, hypothesis test Type B and hypothesis test Type A are evaluated consecutively. The reason is that, if hypothesis test Type B is not rejected, then the constrained hypothesis does not fit significantly worse than the best fitting unconstrained hypothesis. In this way, hypothesis test Type B is a check for constraint misspecification. Severe violations will namely result in rejecting the constraint hypothesis (e.g., $20 < 40 < 30$) and further analyses are redundant. If hypothesis test Type B is not rejected, then hypothesis test Type A is evaluated because hypothesis test Type B cannot distinguish between inequality or equality constraints. In addition, because we are mainly interested in the power of the combination of both hypothesis tests, we introduce a new hypothesis test called Type J. The

power of Type J is the probability of not rejecting hypothesis test Type B times the probability that hypothesis test Type A is rejected given that hypothesis test Type B is not rejected. However, in case of constraint misspecification, we will call it pseudo power. This is because for hypothesis test Type B, power is defined as the probability that the hypothesis is correctly not rejected. Since this is not in accordance with the classical definition of power, we call it pseudo power.

In this article, we make use of the $\bar{F}$ (F-bar) statistic for testing hypothesis test Type A and hypothesis test Type B. The $\bar{F}$ is an adapted version of the well known F statistic often used in ANOVA and linear regression and can deal with order/inequality constraints. The technical details of the $\bar{F}$ statistic are discussed in Appendix B, including a brief historical overview. To calculate the p value of the $\bar{F}$ statistic, we cannot rely on the null distribution of F as in the classical F test. However, we can compute the tail probabilities of the $\bar{F}$ distribution by simulation or via the multivariate normal distribution function. The technical details for computing the p value based on the two approaches are discussed in Appendix C.

Several software routines are available for testing constrained hypotheses using the $\bar{F}$ statistic (hypothesis test Type A and Type B). Ordered means may be evaluated by the software routine 'Confirmatory ANOVA' discussed in Kuiper, Klugkist, and Hoijtink (2010). An extension for linear regression models is available in the `R` package ic.infer or in our own written `R` function `csi.lm()`. The function is available online at `http://github.com/LeonardV/CSI_lm` [2]. Hypothesis test Type A may also be evaluated by the statistical software SAS/STAT® (SAS Institute Inc, 2008) using the PLM procedure.

2.3 Sample-size Tables for order constrained ANOVA

In this section we calculate the sample size according to a power of 0.80 for hypothesis test Type J. We will in particular investigate (a) the gain in power when we impose an increasingly number of correctly specified

[2]Note that the `csi.lm()` function is deprecated. The function has been replaced by the `R` package **restriktor**.

order constraints on the one-way ANOVA model; (b) the pseudo power when some of the means are not in line with the ordered hypothesis.

2.3.1 Correctly specified order constraints

We consider the model $y_i = \mu_1 x_{i1} + \ldots + \mu_k x_{ik} + \epsilon_i,\ i = 1, \ldots, n$, where we assume that the residuals are normally distributed. Data are generated according to this model with uncorrelated independent variables, for $k = 3, \ldots, 8$ groups, and for a variety of real differences among the population means, $f = 0.10$ (small), 0.15, 0.20, 0.25 (medium), 0.30, 0.40 (large), where f is defined according to Cohen (1988, pp. 274–275). We generated 20,000 datasets for $N = 6, \ldots, n$, where n is eventually the sample-size per group at a power of 0.80. The simulated power is simply the proportion of p-values smaller than the predefined significance level. In this study we choose the arbitrary value $\alpha = 0.05$. An extensive description of the simulation procedure is given in Appendix D.

Table 2.1 shows the result of the simulation study in which we investigated the sample size at a power of 0.80 for different effect sizes and an increasing number of order constraints. For example, the first row (n_{k3_0}) presents the sample-sizes per group for an ANOVA with $k = 3$ groups and no constraints. These sample-sizes are equal to those in Cohen (1988) [3]. The second row (n_{k3_1}) shows the sample-sizes per group for $k = 3$ and 1 imposed order constraint, and so on. The values between the parentheses show the relative sample-size reduction. The second column represents the Type I error rates. The values are computed based on the smallest sample size given in the last column (S = 10,000, S is the number of datasets). All results are close to the predefined value of $\alpha = 0.05$, despite the fact that hypothesis test Type J is a composite of hypothesis test Type A and Type B.

The results show that, for any value of f, the sample size decreases with the restrictiveness of the hypothesis. In other words, more information about the means, provided by the order constraints imposed on them, leads to a higher power. For example, in case of a small effect size (f = 0.10) and $k = 4$, the total sample size reduction with 1 constraint is

[3] The unconstrained one-way ANOVA sample-sizes may differ slightly ($\pm$ 1) from the sample-sizes described in Cohen (1988). These differences can completely be attributed to the number of simulation runs.

96 (274-250 = 24, 4 × 24 = 96), with 2 constraints 228 (4 × 57), and with 3 constraints 400 (4 × 100). Noteworthy, within a certain group k and a given number of constraints, the sample size decreases relatively equal across effect sizes. For example, if $k = 4$ and 3 constraints are imposed, the sample size decreases approximately 36%, independent of effect size. In addition, we compared the results of hypothesis test Type J with the results of hypothesis test Type A (not shown here). The results are almost identical and show only some minor fluctuations, which confirms that hypothesis test Type B only plays a significant role when the means are not in line with the imposed order.

2.3.2 Incorrect order of the means

The preceding calculations have all been for sets of means which satisfy the order constraints. Its power (read pseudo power) when the order of the means is not satisfied is also of our concern. In particular we would like to know about the power when the means are not perfectly in line with the ordered hypothesis. In this vein, we focus on the scenario that $k = 4$, $f = 0.10$, 0.25, 0.40 and three order constraints. The two outer means are fixed and only the two middle means are varied. For each value of f five variations are investigated according to the rule $\mu_i\gamma$ (i = 2,3), where $\gamma = 0$, -0.25, -0.50, -0.75, -1, and reflects minor to larger violations.

The results reveal that the power for Hypothesis test Type A (H_{A0} vs. H_{A1}) is largely dominated by the extremes (here the first and last mean). This means that, irrespective of the deviations of the two middle means, the power is almost not affected. The results for hypothesis test Type B (H_{B0} vs. H_{B1}) clearly show that the power to detect mean deviations increases with sample size. We can conclude that the pseudo power for Type J is less affected by minor mean deviations, where large violations may affect the pseudo power severely. This effect becomes more pronounced with larger effect sizes.

Table 2.1: Sample-size table for ANOVA - sample size per group ($k = 3, \ldots, 8$) at a power of 0.80 for Type J ($\alpha = 0.05$), for an increasing number of *correctly specified* order constraints. The value between parentheses is the relative decrease in sample size.

	Type I	$f = 0.10$	0.15	0.20	0.25	0.30	0.35	0.40
n_{k3_0}	.050	323	144	82	53	37	28	22
n_{k3_1}	.050	283 (-12.4%)	126	73	47 (-11.3%)	33	24	19 (-13.6%)
n_{k3_2}	.055	224 (-30.7%)	101	57	37 (-30.2%)	26	19	15 (-31.8%)
n_{k4_0}	.050	274	123	70	45	32	24	19
n_{k4_1}	.047	250 (-08.8%)	112	64	42 (-06.7%)	29	22	17 (-10.5%)
n_{k4_2}	.052	217 (-20.8%)	97	55	36 (-20.0%)	25	19	15 (-21.1%)
n_{k4_3}	.051	174 (-36.5%)	79	44	29 (-35.6%)	20	15	12 (-36.8%)
n_{k5_0}	.050	240	108	61	40	28	21	16
n_{k5_1}	.049	229 (-04.6%)	102	59	37 (-07.5%)	27	20	15 (-06.3%)
n_{k5_2}	.047	204 (-15.0%)	92	52	33 (-17.5%)	24	18	14 (-12.5%)
n_{k5_3}	.049	176 (-26.7%)	78	45	28 (-30.0%)	20	15	12 (-25.0%)
n_{k5_4}	.049	143 (-40.4%)	64	36	23 (-42.5%)	16	12	10 (-37.5%)
n_{k6_0}	.050	215	96	55	36	25	19	15
n_{k6_1}	.046	209 (-02.8%)	93	53	35 (-02.8%)	24	18	14 (-06.7%)
n_{k6_2}	.045	189 (-12.1%)	85	48	31 (-13.9%)	22	17	13 (-13.3%)
n_{k6_3}	.047	169 (-21.4%)	76	43	28 (-22.2%)	20	15	12 (-20.0%)
n_{k6_4}	.051	145 (-32.6%)	65	37	24 (-33.3%)	17	13	10 (-33.3%)
n_{k6_5}	.049	120 (-44.2%)	53	30	20 (-44.4%)	14	11	08 (-46.7%)
n_{k7_0}	.050	196	88	50	33	23	17	14
n_{k7_1}	.046	192 (-02.0%)	87	49	32 (-03.0%)	23	17	13 (-07.1%)
n_{k7_2}	.049	177 (-09.7%)	80	46	30 (-09.1%)	21	16	12 (-14.3%)

n_{k7_3}	.047	161 (-17.6%)	71	41	27 (-18.2%)	19	14	11 (-21.4%)
n_{k7_4}	.046	143 (-27.0%)	65	36	24 (-27.3%)	17	13	10 (-28.6%)
n_{k7_5}	.045	124 (-36.7%)	56	32	20 (-39.4%)	15	11	09 (-35.7%)
n_{k7_6}	.048	103 (-47.4%)	46	26	17 (-48.5%)	12	09	07 (-50.0%)
n_{k8_0}	.050	181	81	46	30	21	16	13
n_{k8_1}	.047	179 (-01.1%)	80	46	30 (-00.0%)	21	16	12 (-07.7%)
n_{k8_2}	.044	167 (-07.7%)	75	43	28 (-06.7%)	20	15	12 (-07.7%)
n_{k8_3}	.048	156 (-13.8%)	69	40	26 (-13.3%)	18	14	11 (-15.4%)
n_{k8_4}	.046	140 (-22.7%)	63	36	23 (-23.3%)	16	12	10 (-23.1%)
n_{k8_5}	.046	126 (-30.4%)	56	32	21 (-30.0%)	15	11	09 (-30.8%)
n_{k8_6}	.047	108 (-40.3%)	49	27	18 (-40.0%)	13	10	08 (-38.5%)
n_{k8_7}	.049	092 (-49.2%)	41	23	15 (-50.0%)	11	08	06 (-53.8%)

2.4 Sample-size tables for Inequality constrained linear regression

In this section we calculate again the sample size according to a power of 0.80 for hypothesis test Type J. But now we impose only an increasing number of correctly specified inequality constraints on the regression coefficients. We consider the model $y_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i,\ i = 1, \ldots, n$, where we assume that the residuals are normally distributed. Data are generated according to this model with correlated independent variables and with fixed and all equal regression coefficients ($\beta_i = 0.10$). This is because in a non-experimental setting, correlated independent variables are the rule rather than the exception. Therefore, we investigate this for the situations where the predictor variables are weakly ($\rho = 0.20$) and strongly ($\rho = 0.60$) correlated. To make a fair comparison with the ANOVA results, we also take $\rho = 0$ into account. Let f^2 be the effect size with f^2 = 0.02 (small), 0.05, 0.08, 0.10 (medium), 0.15, 0.20, 0.25, 0.35 (large), where f is defined according to Cohen (1988, pp. 280–281). All remaining steps are identical to the ANOVA setting. A detailed description of the simulation procedure is given in Appendix E.

The first observations that can be made on the Tables 2.2, 2.3 and 2.4 are that all Type I error values (see second column) are close to the predefined value of $\alpha = 0.05$. The values are computed based on the smallest sample given in the last column. Second, in accordance with the ANOVA results, for any value of f^2, the sample size decreases with the restrictiveness of the hypothesis. Third, the relative decrease is independent of effect size.

Table 2.2 presents the results for $\rho = 0$. When we compare these results with the ANOVA results in Table 2.1 it is clear that imposing inequality constraints (e.g., $\beta_i > 0$) on the regression coefficients leads to a lower power compared to order constraints (e.g., $\mu_1 > \mu_2$). For example, for the case that $k = p = 5$ and 4 constraints, the sample size reduction is approximately 40% and 29%, respectively. Moreover, at the maximum number of inequality constraints (here 5 constraints) the sample-size reduction of about 36% is still less than when the parameters are fully ordered. The results for a more realistic scenario ($\rho = 0.20$) are shown in Table 2.3. The findings at a maximum number of inequality

constraints are comparable with the ANOVA results. For example, the total sample size decrease for $p = 3,5,7$ is approximately 34%, 42% and 47%, respectively.

2.5 Guidelines

If researchers want to use our sample-size tables, then we recommend the following 5 steps:

1. Formulate the hypothesis of interest.

2a. Formulate any expectations about the order of the model parameters in terms of order constraints (i.e. means in an ANOVA setting and regression coefficients in a linear regression setting). For example, the expectation that the first mean (μ_1) is larger than the second (μ_2) and third mean (μ_3) can be formulated in terms of two order constraints, namely $\mu_1 > \mu_2$ and $\mu_1 > \mu_3$.

2b. Formulate any expectations about the sign of the model parameters in terms of inequality constraints. For example, the expectation that three (continuous or dummy) predictor variables are positively associated with the response variable. This can be formulated in terms of three inequality constraints, namely $\beta_1 > 0, \beta_2 > 0$ and $\beta_3 > 0$.

3. Count the number of non-redundant constraints in step 2a and/or 2b and lookup the needed sample-size in one of the sample-size tables.

4. Collect the data.

5. Evaluate the constrained hypothesis.

2.6 Illustrations

To illustrate our method, we consider the CBT and IQ examples. We demonstrate how to use the sample-size tables in practice and we present the `R` code for the **restriktor** package for testing the constrained hypotheses. The results of the analyses are also briefly discussed. The

Table 2.2: Sample-size table for linear regression model - total sample size at a power of 0.80 for Type J ($\alpha = 0.05$) for $p = 3, 5, 7$, $\rho = 0$, and an increasing number of *correctly specified* inequality-constraints. The value between parentheses is the relative decrease in sample size.

	Type I	$f^2 = 0.02$	0.05	0.08	0.10	0.15	0.20	0.25	0.35
n_{p3_0}	.049	550	223	141	114	78	60	49	36
n_{p3_1}	.049	497 (-09.6%)	202	127	103 (-09.6%)	70	54	44	32 (-11.1%)
n_{p3_2}	.048	445 (-19.0%)	180	114	091 (-20.1%)	62	48	39	29 (-19.4%)
n_{p3_3}	.047	391 (-28.9%)	157	100	079 (-30.7%)	55	41	33	25 (-30.5%)
n_{p5_0}	.050	646	263	167	135	93	71	58	44
n_{p5_1}	.050	601 (-06.9%)	243	156	126 (-06.6%)	84	66	54	40 (-09.0%)
n_{p5_2}	.049	557 (-13.7%)	227	142	115 (-14.8%)	79	61	49	37 (-15.9%)
n_{p5_3}	.050	512 (-20.7%)	208	132	107 (-20.7%)	72	55	45	33 (-25.0%)
n_{p5_4}	.049	467 (-27.7%)	190	118	096 (-28.8%)	66	50	41	30 (-31.8%)
n_{p5_5}	.049	424 (-34.3%)	171	108	088 (-34.8%)	59	45	37	27 (-38.6%)
n_{p7_0}	.047	723	297	186	154	104	80	66	50
n_{p7_1}	.048	686 (-05.1%)	279	175	141 (-08.4%)	097	75	61	46 (-08.0%)
n_{p7_2}	.048	644 (-10.9%)	259	164	134 (-12.9%)	091	70	58	43 (-14.0%)
n_{p7_3}	.044	602 (-16.7%)	246	155	125 (-18.8%)	085	65	54	40 (-20.0%)
n_{p7_4}	.050	560 (-22.5%)	226	143	118 (-23.3%)	079	61	50	37 (-26.0%)
n_{p7_5}	.044	520 (-28.0%)	211	134	109 (-29.2%)	074	56	46	34 (-32.0%)
n_{p7_6}	.050	482 (-33.3%)	196	125	100 (-35.0%)	067	52	42	31 (-38.0%)
n_{p7_7}	.050	441 (-39.0%)	180	112	091 (-40.9%)	062	47	38	28 (-44.0%)

Table 2.3: Sample-size table for linear regression model - total sample size at a power of 0.80 for Type J ($\alpha = 0.05$) for $p = 3, 5, 7$, $\rho = 0.20$, and an increasing number of *correctly specified* inequality-constraints. The value between parentheses is the relative decrease in sample size.

	Type I	f^2 =0.02	0.05	0.08	0.10	0.15	0.20	0.25	0.35
n_{p3_0}	.049	549	222	142	114	78	60	49	37
n_{p3_1}	.049	498 (-09.3%)	200	127	103 (-09.6%)	71	53	43	32 (-13.5%)
n_{p3_2}	.048	441 (-19.7%)	177	113	090 (-21.1%)	61	47	38	28 (-24.3%)
n_{p3_3}	.051	370 (-32.6%)	150	094	076 (-33.3%)	52	39	32	24 (-35.1%)
n_{p5_0}	.050	648	263	168	136	93	72	58	44
n_{p5_1}	.049	605 (-06.6%)	247	156	125 (-08.1%)	85	65	53	39 (-11.4%)
n_{p5_2}	.046	563 (-13.1%)	226	143	117 (-14.0%)	79	61	50	37 (-15.9%)
n_{p5_3}	.049	509 (-21.5%)	207	130	105 (-22.8%)	72	55	44	33 (-25.0%)
n_{p5_4}	.053	451 (-30.4%)	180	115	093 (-31.6%)	62	48	39	29 (-34.1%)
n_{p5_5}	.045	387 (-40.3%)	156	098	080 (-41.2%)	54	41	33	24 (-45.4%)
n_{p7_0}	.050	723	296	188	153	105	80	66	50
n_{p7_1}	.049	694 (-04.0%)	282	179	144 (-05.8%)	099	76	62	46 (-08.0%)
n_{p7_2}	.048	651 (-09.9%)	265	169	136 (-11.1%)	092	71	58	43 (-14.0%)
n_{p7_3}	.047	612 (-15.4%)	246	158	126 (-17.6%)	086	66	54	40 (-20.0%)
n_{p7_4}	.049	565 (-21.8%)	229	145	117 (-23.5%)	080	61	50	37 (-26.0%)
n_{p7_5}	.044	514 (-28.9%)	206	132	106 (-30.7%)	072	55	44	33 (-34.0%)
n_{p7_6}	.047	453 (-37.3%)	186	116	094 (-38.5%)	064	49	39	29 (-42.0%)
n_{p7_7}	.049	393 (-45.6%)	159	100	081 (-47.0%)	055	42	34	25 (-50.0%)

Table 2.4: Sample-size table for linear regression model - total sample size at a power of 0.80 for Type J ($\alpha = 0.05$) for $p = 3, 5, 7$, $\rho = 0.60$, and an increasing number of *correctly specified* inequality-constraints. The value between parentheses is the relative decrease in sample size.

	Type I	$f^2 = 0.02$	0.05	0.08	0.10	0.15	0.20	0.25	0.35
n_{p3_0}	.049	549	222	142	114	79	60	49	37
n_{p3_1}	.050	507 (-07.6%)	206	129	105 (-07.8%)	71	54	44	33 (-10.8%)
n_{p3_2}	.052	441 (-19.6%)	181	114	090 (-21.0%)	62	48	39	29 (-21.6%)
n_{p3_3}	.050	334 (-39.1%)	137	086	071 (-37.7%)	48	36	30	21 (-43.2%)
n_{p5_0}	.050	648	263	168	136	93	71	58	44
n_{p5_1}	.045	626 (-03.3%)	254	160	131 (-03.6%)	89	67	55	41 (-06.8%)
n_{p5_2}	.046	575 (-11.2%)	234	149	119 (-12.5%)	81	63	51	38 (-13.6%)
n_{p5_3}	.045	525 (-18.9%)	214	137	109 (-19.8%)	75	57	46	34 (-22.7%)
n_{p5_4}	.053	452 (-30.2%)	185	118	095 (-30.1%)	64	50	40	29 (-34.0%)
n_{p5_5}	.051	344 (-46.9%)	139	088	071 (-47.7%)	48	36	30	22 (-50.0%)
n_{p7_0}	.050	720	297	188	151	104	80	66	50
n_{p7_1}	.045	714 (-00.8%)	291	186	148 (-01.9%)	102	78	64	48 (-04.0%)
n_{p7_2}	.050	675 (-06.2%)	275	175	142 (-05.9%)	096	74	61	45 (-10.0%)
n_{p7_3}	.052	635 (-11.8%)	260	165	134 (-11.2%)	090	70	57	42 (-16.0%)
n_{p7_4}	.046	591 (-17.9%)	240	152	124 (-17.8%)	084	64	53	39 (-22.0%)
n_{p7_5}	.049	531 (-26.5%)	219	137	110 (-27.1%)	076	58	47	35 (-30.0%)
n_{p7_6}	.050	464 (-35.5%)	189	119	095 (-37.0%)	065	49	40	30 (-40.0%)
n_{p7_7}	.045	344 (-52.2%)	139	088	071 (-52.9%)	048	36	30	22 (-56.0%)

output of the `conTest()` function for the ANOVA and regression example is provided in Appendix F and G respectively. The example datasets are available online at `http://github.com/LeonardV/CSI_lm`.

2.6.1 ANOVA

In the introduction, we discussed the following order-constrained hypothesis (step 1):

$$H_{CBT} : \mu_{new\_drug\_CBT} < \mu_{old\_drug\_CBT} < \mu_{no\_drug\_CBT}, \qquad (2.3)$$

where the researchers had clear expectations about the order of the three means. These expectations were translated into two order constraints between the parameters (step 2). The next step, before data collection, is to determine the necessary sample size to obtain a power of say 0.80 (α = 0.05) when the two order constraints are taken into account (step 3). Sample-size tables based on the classical F test show that in case of $k = 3$ and $f = 0.25$ 53 subjects per group (159 subjects in total) are necessary. If the researchers plan to use the $\bar{F}$ test instead of the classical F test, then it can be retrieved from Table 2.1 that with two order constraints 37 subjects (111 subjects in total) are needed (see row n_{k3_2}). That is a total sample-size reduction of about 48 subjects or about 30%. Then, in order to evaluate the order constrained hypothesis, using the `conTest()` function, the following lines of `R` code are required (step 5):

```
R> library(restriktor)
R> data  <- read.csv("depression.csv")
R> model <- "depression ~ -1 + group"
R> fit.anova <- lm(model, data = data)
R> myConstraints1 <- " group1 < group2
                       group2 < group3 "
R> conTest(model = fit.anova, constraints = myConstraints1)
```

In the first line the restriktor package is loaded into `R`. In the second line the observed data are loaded into `R`. The data should be a data frame consisting of two columns. The first column contains the observed depression values, the second column contains the group variable. The third line is the model syntax and it is identical to the model syntax for the `R` function `lm()`. The intercept was removed from the model (`-1`) so that the

regression coefficients correspond to the means as in an one-way ANOVA. An ANOVA model is just a special case of the linear model. Therefore, in the forth line we can make use of the linear model `lm()` function in `R`. The fifth line shows the constraint syntax. The constraints can be specified using a text-based description. In case of a categorical predictor constraints can be specified using the factor-level name (here 1, 2 and 3) preceded by the factor name (here group). The sixth line calls the actual `conTest()` function for testing the order-constrained hypothesis. The arguments to `conTest()` are the fitted unconstrained model (`fit.anova`) and the constraint syntax (`myConstraints1`).

The results (see Appendix F) show that for Hypothesis test Type B the order constrained hypothesis is not rejected in favor of the unconstrained one, $\bar{F}_B = 0.000$, $p = 1.000$ (an $\bar{F}_B$ value of zero implies that the means are completely in line with the imposed order). The results for hypothesis test Type A indicate that the classical null hypothesis is rejected in favor of the constrained hypothesis, $\bar{F}_A = 4.414$, $p = 0.038$. Thus, the results are in line with the expectations of the researchers. Noteworthy, when the order is completely ignored, then the omnibus F test is not significant, $F = 1.718$, $p = 0.168$ (not shown here). This clearly demonstrates that the $\bar{F}$ test has substantially more power than the classical F test.

2.6.2 Multiple regression

The use of the linear regression sample-size tables is comparable with the ANOVA sample-size table. Recall, that in the IQ example, a group of researchers wanted to investigate the relation between the response variable IQ and five predictor variables (step 1), namely social skills (β_1), interest in artistic activities (β_2), use of complicated language patterns (β_3), start walking age (β_4), and start talking age (β_5). Their hypothesis of interest was that the first three predictor variables are positively associated with higher levels of IQ ($\beta_1 > 0$, $\beta_2 > 0$ and $\beta_3 > 0$) and that the last two predictors are negatively associated with IQ ($\beta_4 < 0$, $\beta_5 < 0$) (step 2). Thus a total of five inequality constraints were imposed on the regression coefficients (step 3). Furthermore, the researchers expected a medium effect size ($f^2 = 0.10$) for the omnibus F test and a weak correlation ($\rho = 0.20$) among the predictor variables. All things considered, classical sample-size tables based on the F test reveal that at least 136 subjects are necessary

to obtain a power of 0.80 ($\alpha = 0.05$). However, when the expected positive and negative associations are taken into account, then from Table 2.3 it can be retrieved that by means of imposing five inequality constraints, only 80 subjects are needed to maintain a power of 0.80 (see row n_{p5_5}). That is a substantial sample-size reduction of about 40% or 56 subjects.

The `R` code to evaluate this inequality constrained hypothesis is analogue to the ANOVA example (step 5):

```
R> library(restriktor)
R> data <- read.csv("IQ.csv")
R> model <- "IQ ~ social + artistic + language +
                  walking + talking"
R> fit.lm <- lm(model, data = data)
R> myConstraints2 <- " social   > 0
                       artistic > 0
                       language > 0
                       walking  < 0
                       talking  < 0 "
R> conTest(model fit.lm, constraints = myConstraints2)
```

The results (see Appendix G) show that the inequality constrained hypothesis is not rejected in favor of the unconstrained hypothesis, $\bar{F}_B = 0.211$, $p = 0.847$, and that the null hypothesis is rejected in favor of the constrained hypothesis, $\bar{F}_A = 10.707$, $p = 0.019$. Thus, the results are in line with the expectations of the researchers. The results for the classical F test are again not significant, $F = 2.184$, $p = 0.067$.

2.7 Discussion and Conclusion

In this paper we presented the results of a simulation study in which we studied the gain in power for order/inequality constrained hypotheses. The presented sample-size tables are comparable with the sample-size tables described in Cohen (1988) but with the added benefit that researchers will be able to look up the necessary sample size with a predefined power of 0.80 *and* number of imposed constraints.

We included an increasing number of order constraints in the one-way ANOVA hypothesis test and inequality constraints in the linear regression

hypothesis test. The ANOVA results, for $k = 3, \ldots, 8$ groups, showed that a substantially amount of power can be gained when constraints are included in the hypothesis. Depending on the number of groups involved, a maximum sample-size reduction between 30% and 50% could be gained when the full ordering between the means is taken into account. For $k > 4$ it is questionable whether imposing less than two order constraints is sufficient for the minor gain in power; for $k > 7$ this may be questionable for less than three constraints. Furthermore, we also investigated the effect of constraint misspecification on the power. The results showed that small deviations have only a minor impact on the power.

The linear regression results reveal that, for $p = 3{,}5{,}7$ parameters, the power increases with the restrictiveness of the hypothesis independent of effect size. Again, a substantial power increase between approximately 30% and 50% can be gained when taking a correlation (ρ) of 0.20 between the independent variables into account. These findings are comparable with the ANOVA results, but only apply to the maximum number of constraints. In all other cases, the results showed that an ordering of the parameters leads to a higher power compared to imposing inequality constraints on the parameters. Nevertheless, full ordering of the parameters may be challenging, while imposing inequalities on the parameters may be an easier task. Hence, combining inequality constraints and order constraints may be a solution for applied users.

The current study has some limitations. In the data generating process (DGP) for the ANOVA model, we made some simplifying assumptions: the differences between the means are equally spaced, the sample size is equal in each group, there are no missing data, and the residuals are normally distributed. For the linear regression model, the DGP assumes that the correlations between the independent variables are all equal. In future research, the effects of these assumptions on a possible power drop should be studied. Moreover, we only investigated a limited set of possibilities and extensions for $\alpha = 0.01$ and different power levels are desirable. However, because it is impossible to cover all possibilities, we are currently working on a user-friendly `R` package for constrained hypothesis testing which will include functions for sample-size and power calculations. Despite these limitations, we believe that the presented sample-size tables are a welcome addition to the applied user's toolbox, and may help convincing applied users to incorporate constraints in their hypotheses. Indeed,

notwithstanding the substantial gain in power, constrained hypothesis testing is still largely unknown in the social and behavioral sciences, although the social and behavioral sciences are a good source for ordered tests. For example, in an experimental setting, the parameters of interest (e.g., means) can often be ordered easily. In a non-experimental setting variables such as 'self-esteem', 'depression' or 'anxiety' do not conveniently lend themselves for such ordering, but attributing a positive or a negative sign can often be done without much difficulties.

In conclusion, including prior knowledge into a hypothesis, by means of imposing constraints, results in a substantial gain in power. Researchers who are dealing with inevitable small samples in particular may benefit from this gain. Therefore, we recommend applied users to use these sample-size tables and corresponding software tools to answer their substantive research questions.

Acknowledgments

The first author is a PhD fellow of the research foundation Flanders (FWO) at Ghent university (Belgium) and at Utrecht University (The Netherlands). The second author is supported by a grant from the Netherlands organization for scientific research: NWO-VENI-451-11-008.

A

Hypothesis test Type A and Type B

Consider the standard linear regression model,

$$\begin{aligned} y_i &= \theta_1 x_{i1} + \ldots + \theta_p x_{ip} + \epsilon_i,\ i = 1, \ldots, n. \\ &= \sum_{j=1}^{p} \theta_j x_{ij} + \epsilon_i. \end{aligned} \tag{A.1}$$

Hypothesis test Type A and hypothesis test Type B can be summarized as follows:

Type A:

$$\begin{aligned} H_{A0}: &\quad \mathbf{R}\boldsymbol{\theta} = \mathbf{c} \\ H_{A1}: &\quad \mathbf{R_1}\boldsymbol{\theta} \geq \mathbf{c}\ , \end{aligned} \tag{A.2}$$

Type B:

$$\begin{aligned} H_{B0}: &\quad \mathbf{R_1}\boldsymbol{\theta} \geq \mathbf{c} \\ H_{B1}: &\quad \boldsymbol{\theta} \in \mathbf{R}^p\ . \end{aligned} \tag{A.3}$$

If r is the number of inequality constraints imposed on $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$, and p the number of parameters involved, then let $\mathbf{R}$ be an $r \times p$ matrix with known constants, and $\mathbf{c}$ an $r \times 1$ vector with known constants

(often this vector contains zeros). In an ANOVA, each row of matrix $\mathbf{R}$ is typically a permutation of the p-vector $(-1, 1, 0, \ldots, 0)$ and represents one pairwise constraint. In a linear regression model $\mathbf{R}$ is typically a permutation of the p-vector $(1, 0, \ldots, 0)$ and represents a one parameter constraint. Let $\mathbf{R_1}$ be a submatrix of $\mathbf{R}$ of order $q \times p$, where $q \leq r$. For example, suppose that $p = 4$ and $H_{A0} : \theta_1 = \theta_2 = \theta_3 = \theta_4$ and $H_{A1} : \theta_1 < \theta_2 < \{\theta_3,\ \theta_4\}$ (in H_{A1} no specific order between θ_3 and θ_4 is expected), then

$$\mathbf{R} = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \text{ and } \mathbf{R_1} = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}.$$

Furthermore, at least one of the inequality signs in hypothesis test Type A must be a strict inequality so that the null hypothesis is not included in the constrained hypothesis.

B

The $\bar{F}$ test statistic

In the statistical literature, several approaches have been proposed for testing constrained hypotheses. Silvapulle and Sen (2005) present the state-of-the-art with respect to constrained statistical inference, see also Barlow et al. (1972) and Robertson et al. (1988). In addition to the $\bar{F}$ test statistic, some other test statistics in the framework of linear models are the E-square-bar test ($\bar{E}^2$), the Score test, the Wald test and the likelihood ratio test (LRT) (Gouriéroux, Holly, & Monfort, 1982; Silvapulle & Sen, 2005).

The $\bar{F}$ test can be calculated as follows:

$$\bar{F} = \{RSS(\boldsymbol{\theta}_{H0}) - RSS(\boldsymbol{\theta}_{H1})\}/S^2 \tag{B.1}$$

where $RSS(\boldsymbol{\theta})$ is the residual sum of squares under the hypothesis H and can be computed as follows:

$$RSS(\boldsymbol{\theta}) = \sum_{i=1}^{n} e_i^2, \tag{B.2}$$

where $e_i = (y_i - \hat{y}_i)$ and $\hat{y}_i = \hat{\theta}_1 x_{i1} + \ldots + \hat{\theta}_p x_{ip}$. This term is the main building block for the $\bar{F}$ test statistic.

In the unconstrained setting, the solution to $\hat{\boldsymbol{\theta}}$ can be obtained analytically. In case of constraints, we need to find $\tilde{\boldsymbol{\theta}}$, which is the solution to the constrained optimization problem. There are efficient computer algorithms for this optimization problem. For example, the subroutine solve.QP in the `R` package quadprog (Turlach & Weingessel, 2013) works well in our experience.

The $\bar{F}$ test finds its roots in Kudô (1963) who stated its null distribution, but pioneering steps were made in Bartholomew (1959a, 1959b, 1961b) which discussed the $\bar{\chi}^2$ (chi-square-bar) statistic, for the situation where the covariance matrix $\mathbf{V}$ has the form $\mathbf{V} = \sigma^2\mathbf{W}$ and is completely known. Kudô suggested the $\bar{F}$-statistic in case of k independent normal means with known covariance matrix $\mathbf{W}$ but unknown σ^2, see also Nüesch (1966). Kudô's work was extended by Kudô and Choi (1975) who generalized the result to the case when the covariance matrix is singular. This occurs when the number of imposed inequality constraints on the means exceeds the number of means involved. Yancey, Judge, and Bock (1981) discussed tests of the null hypothesis that a subset of the parameter vector lies in the positive orthant [1] for the special case in which the design matrix in the linear model is orthogonal. It was Wolak (1987) who generalized the results of Yancey et al. to the case of an arbitrary design matrix and general equality and inequality constraints. Silvapulle (1996) elaborated the results of Wolak for the case where the hypotheses are more general than the linear ones.

More recent developments in the context of linear models are for example inequality constrained generalized mixed models, and non-normal models such as logistic and Poisson regression, time series, and proportional hazard models (Davis, 2012). In addition, constrained robust tests have been discussed by Silvapulle (1992a, 1992b), and Van de Schoot, Hoijtink, and Deković (2010) presented a method for testing constrained hypotheses in structural equation models. The problem of constrained tests when there are missing data has been studied by Kim and Taylor (1995); Shi, Zheng, and Guo (2005) and Zheng, Shi, and Guo (2005).

[1] An orthant is any of the n-regions into which n-dimensional Euclidean space is divided by the coordinate planes. For example, in two dimensional space there are four orthants. The positive orthant exists of all vectors with positive coordinates.

C

The null distribution of the $\bar{F}$ test

To compute the tail probabilities of the $\bar{F}$ statistic, we cannot rely on the null distribution of F as in the classical F test. This is because its null distribution has become a mixture of F distributions. Closed form expressions for the mixing weights for $p \leq 4$ can be found in Kudô (1963). The exact computation of the weights for $p > 4$ is a difficult task in general. To deal with this issue, we discuss two suitable approaches. In the first approach, the p value can be computed easily and sufficiently accurately by a simulation approach. Let G denote the cumulative distribution function of the residuals where G is assumed known but σ may be unknown. For example the distribution of the residuals may be normally distributed. Then, the p value for the $\bar{F}$ statistic can be computed by using the following four steps Silvapulle and Sen (2005, pp. 98):

1. Generate independent observations $\{y_{ij} : i = 1, \ldots, n_j, j = 1, \ldots, p\}$ from G.

2. Compute the $\bar{F}$ statistic.

3. Repeat the previous two steps say $B = 100{,}000$ times.

4. Estimate the p value by M/B, where M is the number of times the $\bar{F}$ statistic in the second step exceeded its sample value.

Note that in the first step the observations may be generated from a distribution with any value for the mean and variance because the null distribution of the $\bar{F}$ does not depend on them, see Theorem 3.9.1 in Silvapulle and Sen (2005, pp. 97–98). The advantage of this method is that any error distribution may be used for computing the p value. The disadvantage is an increased computational cost. In the second approach, the p value may be computed economically by first simulating the mixing weights (w_i). The weight w_i is some nonnegative value and is the probability that $\tilde{\boldsymbol{\theta}}$ has exactly i positive elements. The sum of the weights from 0 to q is one. These weights explicitly depend on the covariance matrix of $\hat{\boldsymbol{\theta}}$ (Wolak, 1987). If the constrained set is the nonnegative orthant, then the weights can be computed by using the following five steps (see Silvapulle & Sen, 2005, pp. 79):

1. Generate independent observations $\{y_{ij} : i = 1, \ldots, n_j, j = 1, \ldots, p\}$ from G.
2. Compute $\tilde{\boldsymbol{\theta}}$ subject to $\boldsymbol{\theta} \geq 0$.
3. Count the number of elements of the vector $\tilde{\boldsymbol{\theta}}$ greater than zero.
4. Repeat the previous three steps say $B = 10{,}000$ times.
5. Estimate w_i by the proportion of times $\tilde{\boldsymbol{\theta}}$ has exactly i positive elements, $i = 0, \ldots, q$.

In addition, if the residuals are normally distributed, then the weights can be computed by using the multivariate normal probability distribution function. This method is implemented in the `ic.weight()` function in the `R` package ic.infer (Grömping, 2010).

Then, the p value for hypothesis test Type A can be computed as follows Silvapulle and Sen (2005, pp. 99):

$$\Pr(\bar{F}_A \geq \bar{f}_{A_{obs}}) = \sum_{i=0}^{q} w_i(H_0, H_1)\ \Pr[(r-q+i)F_{r-q+i,\nu} \geq \bar{f}_{A_{obs}}], \quad \text{(C.1)}$$

where ν is the error degrees of freedom and the $\bar{f}_{obs}$ is the sample value of the $\bar{F}$. For hypothesis test Type B, with only order/inequality constraints, the p value is computed as Silvapulle and Sen (2005, pp. 100):

$$\Pr(\bar{F}_B \geq \bar{f}_{B_{obs}}) = \sum_{i=0}^{q} w_i(H_0, H_1) \ \Pr[iF_{i,\nu} \geq \bar{f}_{B_{obs}}]. \qquad \text{(C.2)}$$

D

Simulation 1 - correctly specified order constraints

In a one-way ANOVA, the populations differ only in their means. Let $\boldsymbol{\theta} = (\mu_1, \mu_2, \ldots, \mu_k)$ and let f be a measure of the true deviation from the null hypothesis, where f is defined according to Cohen (1988, pp. 274–275). Next, we discuss our six step simulation procedure.

In step 1, data are generated according to the model specified in Equation A.1 with uncorrelated independent variables, for $k = 3, \ldots, 8$ groups and for a variety of real differences among the population means, $f = 0.10$ (small), 0.15, 0.20, 0.25 (medium), 0.30, 0.40 (large). Let the differences between the means, d, be equally spaced. Then d is defined as $d = \frac{2f\sqrt{k}}{\sqrt{\sum_{i=1}^{k}(2i-1-k)^2}}$ under the restriction that $\sum_{i=1}^{k} \mu_i = 0$ and $\sigma = 1$. The smallest mean, μ_1, is determined by $\mu_1 = \frac{-(k-1)d}{2}$. For example, if $k = 4$ and the effect size $f = 0.25$, then $d = \sqrt{\frac{1}{20}}$ and $\mu_1 = -0.335$. Then, $\mu_2 = \mu_1 + d$, $\mu_3 = \mu_1 + 2d$ and $\mu_4 = \mu_1 + 3d$.

In step 2, we generate $S = 20,000$ datasets according to the data generating process described in step 1 for $N = 6, \ldots, n$, where n is eventually the sample size per group at a power of 0.80.

In step 3, we fit the equality-constrained model (H_{A0}), the order-constrained model (H_{A1}) and the two-sided unconstrained model (H_{B1}) and calculate for each model the RSS_H. The imposed order constraints are of the form H: $\mu_t - \mu_s \geq 0$.

In step 4, we calculate the $\bar{F}$ values for hypothesis test Type A and Type B according to Equation B.1.

Then, in step 5, we compute the p-value for hypothesis test Type A and hypothesis test Type B. This is done according to Equation C.1 for hypothesis test Type A and according to Equations C.2 for hypothesis test Type B, which are provided in Appendix C.

Finally in step 6, we calculate the power for hypothesis tests Type A, Type B, and Type J. The power is simply the proportion of p-values smaller than the predefined significance level. In this study we choose the arbitrary value $\alpha = 0.05$. The conditional power is computed by $\hat{P}(\bar{b}) \times \hat{P}(a|\bar{b})$, where $\hat{P}(a)$ is the proportion of significant results for hypothesis test Type A, and $\hat{P}(\bar{b})$ is the proportion of non-significant results for hypothesis test Type B.

E

Simulation 2 - correctly specified inequality constraints

In a linear regression analysis, let $\boldsymbol{\theta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ and let $f^2 = \frac{R^2}{1-R^2}$, where R^2 is the determination coefficient. Again, a six step simulation procedure is used, similar as for the ANOVA setting.

In step 1, data are generated according to Equation A.1 with fixed and all equal parameters ($\beta_i = 0.10$). Let f^2 indicate the effect size with f^2 = 0.02 (small), 0.05, 0.08, 0.10 (medium), 0.15, 0.20, 0.25, 0.35 (large). Since we hold the parameters fixed, generating data for a predefined R^2 boils down to determining σ^2, where $\sigma^2 = (\boldsymbol{\theta}^\top \boldsymbol{\Sigma_X} \boldsymbol{\theta})(1 - R^2)/R^2$ and $\boldsymbol{\Sigma_X}$ is the covariance-matrix for the covariances between the independent variables. We take this latter into account because in a non-experimental setting, correlated independent variables are the rule rather than the exception. Therefore, we investigate this for the situations where $\boldsymbol{\Sigma_X}$ is a compound symmetry matrix with ones on the diagonal and values of ρ (ρ = 0, 0.20 and 0.60) elsewhere. We take the value $\rho = 0$ into account to make a fair comparison with the ANOVA model. Furthermore, in this

study we will limit ourselves to $p = 3$, 5, and 7 variables.

In step 2, we generate $S = 20,000$ datasets according to the data generating process described in step 1 for $N = 6, \ldots, n$, where n is the total sample size at a power of 0.80.

Step 3 corresponds to the ANOVA setting with the exception that we impose an increasing number of correctly specified inequality constraints of the form H: $\beta_i \geq 0$ on the model.

Step 4, 5 and 6 are again identical to the ANOVA setting.

F

Output of the conTest() function for the CBT example

```
Restriktor: restricted hypothesis tests ( 108 residual degrees of freedom ):


Multiple R-squared remains 0.046

Constraint matrix:
   factor(group)1 factor(group)2 factor(group)3    op rhs active
1:             -1              1              0    >=   0     no
2:              0             -1              1    >=   0     no

Overview of all available hypothesis tests:

Global test: H0: all parameters are restricted to be equal (==)
         vs. HA: at least one inequality restriction is strictly true (>)
       Test statistic: 4.4144,   p-value: 0.03814

Type A test: H0: all restrictions are equalities (==)
         vs. HA: at least one inequality restriction is strictly true (>)
       Test statistic: 4.4144,   p-value: 0.03814
```

```
Type B test: H0: all restrictions hold in the population
         vs. HA: at least one restriction is violated
       Test statistic: 0.0000,   p-value: 1

Type C test: H0: at least one restriction is false or active (==)
         vs. HA: all restrictions are strictly true (>)
       Test statistic: 0.9968,   p-value: 0.1605

Note: Type C test is based on a t-distribution (one-sided),
      all other tests are based on a mixture of F-distributions.
```

G

Output of the conTest() function for the IQ example

```
Restriktor: restricted hypothesis tests ( 65 residual degrees of freedom ):


Multiple R-squared reduced from 0.144 to 0.141

Constraint matrix:
   (Intercept) social artistic language walking talking    op rhs active
1:           0      1        0        0       0       0    >=   0     no
2:           0      0        1        0       0       0    >=   0     no
3:           0      0        0        1       0       0    >=   0     no
4:           0      0        0        0      -1       0    >=   0     no
5:           0      0        0        0       0      -1    >=   0    yes

Overview of all available hypothesis tests:

Global test: H0: all parameters are restricted to be equal (==)
         vs. HA: at least one inequality restriction is strictly true (>)
       Test statistic: 10.7071,   p-value: 0.01934
```

```
Type A test: H0: all restrictions are equalities (==)
         vs. HA: at least one inequality restriction is strictly true (>)
       Test statistic: 10.7071,   p-value: 0.01934

Type B test: H0: all restrictions hold in the population
         vs. HA: at least one restriction is violated
       Test statistic: 0.2109,   p-value: 0.8472

Type C test: H0: at least one restriction is false or active (==)
         vs. HA: all restrictions are strictly true (>)
       Test statistic: -0.4593,   p-value: 0.6762

Note: Type C test is based on a t-distribution (one-sided),
      all other tests are based on a mixture of F-distributions.
```

References

Barlow, R., Bartholomew, D., Bremner, H., & Brunk, H. (1972). *Statistical inference under order restrictions*. New York: Wiley.

Bartholomew, D. (1959a). A test of homogeneity for ordered alternatives. *Biometrika*, *46*, 36–48.

Bartholomew, D. (1959b). A test of homogeneity for ordered alternatives. II. *Biometrika*, *46*, 328–335.

Bartholomew, D. (1961a). Ordered tests in the analysis of variance. *Biometrika*, *48*(3/4), 325–332. doi: doi:10.2307/2332754

Bartholomew, D. (1961b). A test of homogeneity of means under restricted alternatives. *Journal of the royal statistical society. Series B (Methodological)*, *23*(2), 239–281.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Davis, K. (2012). Constrained statistical inference: A hybrid of statistical theory, projective geometry and applied optimization techniques. *Progress in Applied Mathematics*, *4*(2), 167–181.

Gouriéroux, C., Holly, A., & Monfort, A. (1982). Likelihood ratio test, wald test, and kuhn-tucker test in linear models with inequality constraints on the regression parameters. *Econometrica*, *50*, 63–80. doi: doi:10.2307/1912529

Grömping, U. (2010). Inference with linear equality and inequality constraints using R: The package ic.infer. *Journal of statistical software*, *33*, 1–31. doi: doi:10.18637/jss.v033.i10

Hoijtink, H. (2012). *Informative Hypotheses: Theory and Practice for Behavioral and Social Scientists*. Boca Raton, FL: Taylor & Francis.

Kim, D., & Taylor, J. (1995). The restricted EM algorithm for maximum likelihood estimation under linear restrictions on the parameters. *Journal of the American Statistical Association*, *90*(430), 708–716.

Kudô, A. (1963). A multivariate analogue of the one-sided test. *Biometrika*, *50*(3/4), 403–418. doi: doi:10.2307/2333909

Kudô, A., & Choi, J. (1975). A generalized multivariate analogue of the one sided test. *Memoirs of the faculty of science*, *29*, 303–328.

Kuiper, R., & Hoijtink, H. (2010). Comparisons of means using exploratory and confirmatory approaches. *Psychological Methods*, *15*(1), 69–86. doi: doi:10.1037/a0018720

Kuiper, R., Klugkist, I., & Hoijtink, H. (2010). A Fortran 90 Program for Confirmatory Analysis of Variance. *Journal of Statistical Software*, *34*, 1–31.

Kuiper, R., Nederhoff, T., & Klugkist, I. (2011, October). *Performance and robustness of confirmatory approaches.* (The paper is available at http://informative-hypotheses.sites.uu.nl/wp-content/uploads/sites/23/2015/04/Kuiper-Nederhoff-and-Klugkist.pdf)

Nüesch, P. (1966). On the problem of testing location in multivariate populations for restricted alternatives. *The Annals of Mathematical Statistics*, *37*, 113–9.

Perlman, M. (1969). One-sided testing problems in multivariate analysis. *The Annals of Mathematical Statistics*, *40*(2), 549–567. doi: doi:10.1214/aoms/1177697723

R Development Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. (ISBN 3-900051-07-0)

Robertson, T., Wright, F. T., & Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. New York: Wiley.

SAS Institute Inc. (2008). Sas/stat® 9.2 user's guide [Computer software manual]. Cary, NC: SAS Institute Inc.

Shi, N., Zheng, S., & Guo, J. (2005). The restricted EM algorithm under inequality restrictions on the parameters. *Journal of Multivariate Analysis*, *92*(1), 53–76.

Silvapulle, M. (1992a). Robust tests of inequality constraints and one-sided hypotheses in the linear model. *Biometrika*, *79*(3), 621–630. doi: doi:10.2307/2336793

Silvapulle, M. (1992b). Robust wald-type tests of one-sided hypotheses in the linear model. *Journal of the American Statistical Association*, *87*(417), 156–161. doi: doi:10.2307/2290464

Silvapulle, M. (1996). On an F-type statistic for testing one-sided hypotheses and computation of chi-bar-squared weights. *Statistics & probability letters*, *28*, 137–141.

Silvapulle, M., & Sen, P. (2005). *Constrained statistical inference: Order, inequality, and shape restrictions.* Hoboken, NJ: Wiley.

Turlach, B., & Weingessel, A. (2013). quadprog: Functions to solve quadratic programming problems (version 1.5-5). [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=quadprog`

Van de Schoot, R., Hoijtink, H., & Deković, M. (2010). Testing inequality constrained hypotheses in SEM models. *Structural Equation Modeling*, *17*, 443–463. doi: doi:10.1080/10705511.2010.489010

Van de Schoot, R., & Strohmeier, D. (2011). Testing informative hy-

potheses in SEM increases power: An illustration contrasting classical hypothesis testing with a parametric bootstrap approach. *International Journal of Behavioral Development*, *35*, 180–190. doi: doi:10.1177/0165025410397432

Wolak, F. (1987). An exact test for multiple inequality and equality constraints in the linear regression model. *Journal of the American statistical association*, *82*(399), 782–793. doi: doi:10.1080/01621459.1987.10478499

Wolak, F. (1989). Testing inequality constraints in linear econometric models. *Journal of Econometrics*, *41*(2), 205–235. doi: doi:10.1016/0304-4076(89)90094-8

Yancey, T., Judge, G., & Bock, M. (1981). Testing multiple equality and inequality hypothesis is economics. *Economics Letters*, *7*, 249–255.

Zheng, S., Shi, N., & Guo, J. (2005). The restricted EM algorithm under linear inequalities in a linear model with missing data. *Science in China Series A – Mathematics*, *48*(6), 819–828.

3

Comparing inequality-constrained robust and non-robust regression estimation methods for one-sided hypotheses

In many situations, researchers have specific expectations about the order of the parameters in their statistical model. For example, a researcher might expect that the regression coefficients follow a simple order (e.g., $\theta_1 \leq \theta_2 \leq \theta_3$). Contaminated data, such as extreme observations in the response and the predictor space are ubiquitous in research. Both may have a great negative impact on the least squares estimator resulting in bias and loss in power. Robust estimation of the linear model, where extreme observations are down-weighted to have less influence on the parameter estimates is a powerful alternative to least squares. The result is a robustly estimated constrained linear model. We investigate by means of a simulation study the performance of inequality-constrained (IC) regression OLS-, M- and MM-estimators in terms of their mean squared error

(MSE). Moreover, we investigate the size and power of one-sided robust and non-robust hypothesis tests (Wald, score and likelihood ratio). The results show that IC MM-estimation produces the most precise estimates, while the M- and OLS-estimates are negatively affected for higher levels of contamination. The power of the OLS- and M-tests fails dramatically, while the power of the robust MM-tests remains adequate. An empirical example about child and parental adjustment following a pediatric burn event illustrates the application of these robust tests and shows that ignoring extreme observations in the analysis may result in spurious conclusions regarding the direction of the effects. Therefore, we advise robust techniques if the data are potentially contaminated.

3.1 Introduction

Small samples and extreme observations are often encountered in research. The easiest way to overcome the problem of too small samples and related lack of power is to find a way to increase the sample size. Unfortunately, this is often impossible due to limited resources (e.g., in expensive fMRI studies), ethical issues (e.g., in case of vulnerable groups) or small populations (e.g., in clinical trials). Research into the psychological consequences of pediatric burns concerns a typical example of a research field in which problems with regard to sample size arise. Burn centers often comprise small units, resulting in the need for prolonged multi-center studies in order to obtain a sufficient sample size. Consequently, many research questions remain unanswered.

Inequality-constrained (IC) hypothesis testing (Barlow, Bartholomew, Bremner, & Brunk, 1972; Robertson, Wright, & Dykstra, 1988; Silvapulle & Sen, 2005), also known as 'informative hypothesis testing' (Hoijtink, 2012) might be an easy solution. Researchers often have a-priori expectations about the sign or ordering of the parameters in their statistical model. Most researchers are familiar with this fact in the context of the one-sided t-test, where one mean is restricted to be larger or smaller than a fixed value (e.g., $\mu_1 \geq 0$) or another mean (e.g., $\mu_1 \leq \mu_2$). This readily extends to the setting where more than one constraint can be imposed on the statistical parameters. For example, a researcher might expect that a subset of $\boldsymbol{\theta}$, e.g., θ_1, θ_2, where $\boldsymbol{\theta}$ is a vector containing regression

coefficients, is larger than zero $H : \{\theta_1, \theta_2\} \geq \mathbf{0}$ (or any other constant value). In other words, researchers may have clear expectations about the sign (positive or negative) of the parameters in their statistical models. Alternatively, researchers may have clear a-priori expectations about the order of the parameters in a statistical model. For example in a regression model, they may expect that the regression coefficients are subject to order constraints, e.g., $\theta_1 \leq \theta_2 \leq \theta_3$. A major advantage of including inequality constraints in the hypothesis is that power can be gained. This has been shown repeatedly (Barlow et al., 1972; Bartholomew, 1961a, 1961b; Kuiper & Hoijtink, 2010; Meyer & Wang, 2012; Perlman, 1969; Robertson et al., 1988; Rosen & Davidov, 2012; Vanbrabant, Van de Schoot, & Rosseel, 2015) for the linear model using OLS estimators and normal distributed data. In particular, Vanbrabant et al. (2015) have shown that a sample size reduction up to 50% can be achieved if a maximum number of constraints is imposed on the regression coefficients.

Unfortunately, data collected from a wide range of applications often contain irregularities that deviate from the majority of the data (Hampel, 1973; Maronna, Martin, & Yohai, 2006), such as response outliers and bad-leverage points in a regression setting. Response outliers are defined as extreme observations in the response space and bad-leverage points are defined as observations that are extreme in both the response and predictor space. Both may largely affect the OLS-estimator, resulting in biased estimates and a decline in power (Schrader & Hettmansperger, 1980; Silvapulle, 1992a, 1992b). To deal with these issues, various robust estimators have been proposed. Among these are the commonly used M-estimators (Huber, 1973), S-estimators (Rousseeuw & Yohai, 1984) and MM-estimators (Yohai, 1987). Robust estimators achieve their robustness by modifying the loss function, making it less increasing than the squared loss in OLS. Robustness of these estimators can be investigated via their breakdown point (BDP) while performance can be studied by their relative efficiency. Simply put, the BDP of a parameter estimate $\hat{\theta}_j$ is the largest proportion of irregularities that the data may contain such that $\hat{\theta}_j$ still gives some information about θ_j (Maronna et al., 2006). Thus, the higher the BDP the more robust is the estimator. The non-robust OLS has a zero BDP, which means that a single data-point can already distort the OLS estimator. The relative efficiency of an estimator is the ratio of its variance compared to that of the optimal (smallest

variance) estimator. Since the OLS estimator is, under the Gauss-Markov assumptions a best (smallest variance) linear unbiased estimator, robust estimators are frequently compared to OLS, in the ideal case of normally distributed errors. M-estimators can attain a high relative efficiency (over 95%) and can handle response outliers, but unfortunately they can still be unduly influenced by even a single extreme bad-leverage point and therefore have a zero BDP as well. S-estimators have a high BDP of 50%, but they can only attain a relative efficiency up to 33%. MM-regression estimators combine the strengths of M- and S-estimators, so that MM-estimators can simultaneously achieve a high BDP and a high efficiency. In MM-estimation, the initial regression coefficients and final scale estimate are computed by an S-estimator; this determines the BDP. The final estimator of the regression coefficients is an M-estimator with fixed scale equal to the S-scale estimate. This MM-estimator inherits the BDP from the S-estimator in the first step while the M-estimator in the second step determines its relative efficiency (Yohai, 1987).

The natural result of combining both fields of constrained statistical inference and robust estimation results in robust constrained inference for the linear model. Robust Wald, score and likelihood ratio type (LRT) tests for one-side hypotheses based on IC M-estimators have been introduced by Silvapulle (Silvapulle, 1992a, 1992b, 1996) and Silvapulle and Silvapulle (Silvapulle & Silvapulle, 1995). These authors showed in a small simulation study for $n = 18$, two inequality constraints and several error distributions that substantial power can be gained. However, as discussed above, bad-leverage points may have a negative impact on the BDP of the M-estimator. Therefore, we extend their research to IC MM-estimators. The objective of the current paper is to investigate by means of a simulation the performance of IC OLS-, M- and MM-estimators and corresponding tests when the data are contaminated with outliers and bad-leverage points. We show by means of an empirical example about a pediatric burn event that ignoring extreme observations in the analysis may result in spurious results regarding the direction of the effects.

The remainder of this article is organized as follows. First, we describe the linear model using OLS-, M- and MM-estimators and IC hypothesis tests. Second, we describe three non-robust test-statistics and three robust test-statistics that can deal with inequality constraints and we discuss their null-distributions. Third, we present the results of our simulation

study. Our analysis shows that MM-estimation produces the most accurate estimates, while the estimates for OLS- and M-estimation are severely negatively affected for higher levels of contamination. The power fails dramatically for OLS- and M-tests, while the power for MM-tests remains adequate. In addition, for all tests, incorporating order/inequality constraints yield a substantial improvement of the size and power. Next, we present an empirical data example about pediatric burn events. The example shows that ignoring extreme observations in the analysis may result in spurious conclusions regarding the direction of the effects. Therefore, we advise robust techniques if the data are potentially contaminated. Finally, we present a conclusion of our research.

3.2 Linear model and inequality constrained hypotheses

Consider the standard linear regression model,

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\theta} + \epsilon_i, \ i = 1, \ldots, n, \tag{3.1}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$ is the parameter vector of interest, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ are vectors of covariates, and $\epsilon_i = (\epsilon_1, \ldots, \epsilon_i)^T$ are the random errors. In case the model contains an intercept, we set $x_{i1} \equiv 1$ and the vector of regression coefficients can be split in an intercept component $\alpha = \theta_1$ and a slope component $\boldsymbol{\beta} = (\theta_2, \ldots, \theta_p)^T$. Moreover, in this case we write $\boldsymbol{x}_i = (1, \boldsymbol{z}_i^T)^T$ and we assume that the covariates are centered, i.e. $\sum_{i=1}^n z_{ij} = 0$ for $j = 1, \ldots, p-1$. In absence of an intercept, we set $\boldsymbol{\beta} = \boldsymbol{\theta}$. We consider the following methods to estimate $\boldsymbol{\theta}$.

3.2.1 OLS-estimation

Unconstrained OLS estimates $\hat{\boldsymbol{\theta}}^L$ are obtained as the solution which minimizes

$$\sum_{i=1}^{n} \rho\left(e_i\right) \tag{3.2}$$

over all $\boldsymbol{\theta} \in \mathbb{R}^p$, where the loss function equals $\rho(e_i) = e_i(\boldsymbol{\theta})^2$ and $e_i(\boldsymbol{\theta}) = y_i - x_i^T \boldsymbol{\theta}$.

3.2.2 M-estimation

The letter M indicates that it is an extension of the maximum likelihood estimation method. The unconstrained M-estimator $\hat{\boldsymbol{\theta}}^M$ is obtained as the solution which minimizes the loss function

$$\sum_{i=1}^{n} \rho\left(\frac{e_i}{\hat{\sigma}}\right) \tag{3.3}$$

over all $\boldsymbol{\theta} \in \mathbb{R}^p$. In contrast to OLS, the estimation of $\hat{\boldsymbol{\theta}}^M$ is dependent on an initial scale estimate $\hat{\sigma}$. To obtain a robust solution, a robust scale estimator needs to be used to obtain $\hat{\sigma}$. Typically, the MAD (Median Absolute Deviation) of the residuals with respect to an initial estimator (OLS) is used. For the loss function ρ, a common choice is the redescending Tukey biweight (bisquare) family of loss functions, given by

$$\rho(\, e_i; c\,) = \left\{ \begin{array}{ll} 1 - (1 - (e_i/c)^2)^3 & \text{if } |e_i| \leq c \\ 1 & \text{if } |e_i| \geq c \end{array} \right\}, \tag{3.4}$$

with derivative $\rho'(\, e_i; c\,) = 6\psi(\, e_i; c\,)/c^2$ where,

$$\psi(\, e_i; c\,) = e_i \left(1 - \left(e_i/c\right)^2\right)^2 \times I_{\{|e_i| \leq c\}}, \tag{3.5}$$

The indicator function I equals 1 if the expression inside the curly brackets is true and 0 otherwise. Setting the tuning constant $c > 0$ equal to $c = 4.685$ yields an M-regression estimator with 95% efficiency at the central model with normal errors. It is important to note that equation 3.3 can be written as a weighted least-squares problem with weights equal to $w_i = w(e_i) = \psi(e_i)/e_i$ and can be solved using iteratively reweighted least-squares (IRLS).

3.2.3 MM-estimation

MM-estimators are based on two loss functions ρ_1 and ρ_2 which determine the BDP and the efficiency of the estimator respectively. Both loss functions are taken from the Tukey biweight family of loss functions which yields an MM-estimator that is robust to both response and (bad)-leverage

points. Similarly as for the M-estimator, the MM-estimator $\hat{\boldsymbol{\theta}}^{MM}$ is obtained as the solution which minimizes the loss function

$$\sum_{i=1}^{n} \rho_2 \left(\frac{e_i}{\hat{\sigma}^S} \right) \tag{3.6}$$

over all $\boldsymbol{\theta} \in \mathbb{R}^p$. Here, $\hat{\sigma}^S$ is a scale S-estimate. A scale S-estimator is defined as the solution $\hat{\sigma}^S$ which minimizes the M-scale $\hat{\sigma}^M(\boldsymbol{\theta})$ over all $\boldsymbol{\theta} \in \mathbb{R}^p$, where for any $\boldsymbol{\theta} \in \mathbb{R}^p$ the corresponding M-scale $\hat{\sigma}^M(\boldsymbol{\theta})$ is defined implicitly by the equation

$$\frac{1}{n} \sum_{i=1}^{n} \rho_1 \left(\frac{e_i}{\hat{\sigma}^M(\boldsymbol{\theta})} \right) = \text{b}. \tag{3.7}$$

The constant b is usually chosen to obtain a consistent estimator in case of normal errors. The associated S-regression estimator is the solution $\hat{\boldsymbol{\theta}}^S$ which minimizes $\hat{\sigma}^M(\boldsymbol{\theta})$ over all $\boldsymbol{\theta} \in \mathbb{R}^p$, that is $\hat{\sigma}^S = \hat{\sigma}^M(\hat{\boldsymbol{\theta}}^S)$. This S-regression estimate is used as initial value to calculate $\hat{\boldsymbol{\theta}}^{MM}$ using an IRLS procedure to minimize the loss function in 3.6. The constant c in ρ_1 equals 1.548 to obtain an S/MM-estimator with a BDP of 50% while the constant c in ρ_2 equals 4.685 to obtain an MM-regression estimator with 95% efficiency in case of normal errors.

3.2.4 Inequality constrained hypotheses

Let the null and alternative hypotheses be

$$H_0 : \boldsymbol{\theta} \in \mathcal{M} \quad \text{and} \quad H_1 : \boldsymbol{\theta} \in \mathcal{C}, \boldsymbol{\theta} \notin \mathcal{M}, \tag{3.8}$$

where $\mathcal{M}$ is a subspace in $\mathcal{C}$, and $\mathcal{M}$ and $\mathcal{C}$ are subsets of the p-dimensional Euclidean space $\mathbb{R}^p$. If we only consider linear hypotheses, then the null- and alternative hypotheses can be written in the more familiar form $H_0 : \boldsymbol{R\theta} = \mathbf{0}$ and $H_1 : \boldsymbol{R_1\theta} \geq \mathbf{0}$, respectively. If r is the number of inequality constraints imposed on $\boldsymbol{\theta}$, and p the number of parameters involved, then let $\boldsymbol{R}$ be an $r \times p$ matrix with known constants. Let $\boldsymbol{R_1}$ be a submatrix of $\boldsymbol{R}$ of order $q \times p$, where $q \leq r$. Note that at least one of the inequality signs must be a strict inequality so that the null hypothesis is not included in the alternative hypothesis. For a detailed discussion of this type of hypotheses see Silvapulle and Sen (2005).

When the error distribution is asymmetric, the intercept and the center of the error distribution are confounded (Silvapulle, 1992b). Therefore, we restrict ourselves to the case where the hypotheses involve the slope component $\boldsymbol{\beta}$ of $\boldsymbol{\theta}$ only, i.e. $H_0 : \boldsymbol{R\beta} = \mathbf{0}$ against $H_1 : \boldsymbol{R_1\beta} \geq \mathbf{0}$. This restriction should not pose any issues since in most practical situations we only have prior knowledge about the signs of the slope components $\boldsymbol{\beta}$, while the sign of the intercept $\theta_1 = \alpha$ can be changed arbitrarily by shifting the response $\mathbf{y}$.

Calculating least squares estimates $\tilde{\boldsymbol{\theta}}^L$ of the linear regression coefficients under constraints is a well-studied problem (Nocedal & Wright, 2006) and routines are widely available in software, for example in the `R` (R Development Core Team, 2016) package **quadprog** (Turlach & Weingessel, 2013). To calculate constrained M-estimates $\tilde{\boldsymbol{\theta}}^M$ and MM-estimates $\tilde{\boldsymbol{\theta}}^{MM}$, we exploit that in absence of constraints both estimators can be calculated by IRLS. To incorporate the constraints, we replace the IRLS steps by iteratively reweighted constrained least squares optimization steps (IRCLS). In Algorithm 1 we show the core steps of the IRCLS algorithm in pseudo code.

3.3 Test-statistics and null-distributions

First, we describe a non-robust F test (Kudô, 1963), a likelihood ratio (LR) test and a score test (Silvapulle & Sen, 2005) based on IC OLS-estimators. Then, we describe robust counterparts for these tests. We consider a robust Wald test (Silvapulle, 1992b), a likelihood ratio type (LRT) test (Silvapulle, 1992a) and a score test (Silvapulle, 1996) based on IC M- and MM-estimators. Finally, we discuss the null-distributions of these test-statistics.

3.3.1 Non-robust F, LR and score test-statistic

Denote the OLS-estimates for the null, unconstrained and IC model by $\hat{\boldsymbol{\beta}}_0^L$, $\hat{\boldsymbol{\beta}}^L$, and $\tilde{\boldsymbol{\beta}}^L$, respectively. Let us denote the IC F test by $\overline{F}$, the LR test by $\overline{LR}$ and the score test by $\overline{S}$. The bar in the notation indicates that we use the IC counterpart of the corresponding unconstrained test-

statistics. Then, the $\overline{F}$ test-statistic is given by

$$\begin{aligned}\overline{F} = \inf_b\{(\hat{\boldsymbol{\beta}}^L - \boldsymbol{b})^T\widehat{\boldsymbol{W}}^{-1}(\hat{\boldsymbol{\beta}}^L - \boldsymbol{b}) : \boldsymbol{Rb} = \boldsymbol{0}\} - \\ \inf_b\{(\hat{\boldsymbol{\beta}}^L - \boldsymbol{b})^T\widehat{\boldsymbol{W}}^{-1}(\hat{\boldsymbol{\beta}}^L - \boldsymbol{b}) : \boldsymbol{R_1 b} \geq \boldsymbol{0}\},\end{aligned} \tag{3.9}$$

with $\widehat{\boldsymbol{W}} = \hat{\sigma}^2(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}$ where the matrix $\boldsymbol{Z}$ contains the vectors $\boldsymbol{z}_i$ as its rows. Moreover, $\hat{\sigma}^2$ is a consistent estimate of the asymptotic variance σ^2 of the estimator $\hat{\boldsymbol{\beta}}^L$.

The LR test-statistic is given by

$$\overline{LR} = -2[\inf_b\{L(\boldsymbol{b}) : \boldsymbol{Rb} = \boldsymbol{0}\} - \inf_b\{L(\boldsymbol{b}) : \boldsymbol{R_1 b} \geq \boldsymbol{0}\}], \tag{3.10}$$

where $L(\boldsymbol{b}) = \sum_i^n e_i^2$ with $e_i = y_i - \hat{\alpha}^L - \boldsymbol{z}_i^T\boldsymbol{b}$.

The score test-statistic is given by

$$\begin{aligned}\overline{S} = \inf_b\{(S(\hat{\boldsymbol{\beta}}^L) - \boldsymbol{b})^T\widehat{\boldsymbol{W}}_0^{-1}(S(\hat{\boldsymbol{\beta}}^L) - \boldsymbol{b}) : \boldsymbol{Rb} = \boldsymbol{0}\} - \\ \inf_b\{(S(\hat{\boldsymbol{\beta}}^L) - \boldsymbol{b})^T\widehat{\boldsymbol{W}}_0^{-1}(S(\hat{\boldsymbol{\beta}}^L) - \boldsymbol{b}) : \boldsymbol{R_1 b} \geq \boldsymbol{0}\},\end{aligned} \tag{3.11}$$

where $S(\hat{\boldsymbol{\beta}}^L)$ is the vector of the unconstrained scores vector and $\widehat{\boldsymbol{W}}_0 = \hat{\sigma}_0^{-2}\boldsymbol{Z}^T\boldsymbol{Z}$ and $\hat{\sigma}_0^2$ consistently estimates the asymptotic variance σ_0^2 of the estimator $\hat{\boldsymbol{\beta}}_0^L$.

3.3.2 Robust Wald, score and LRT test-statistic

Let us denote the robust Wald test by $\overline{RW}$, the robust score test by $\overline{RS}$ and the robust LRT test by $\overline{RF}$. Denote the null, unconstrained and inequality constrained robust estimates by $\hat{\boldsymbol{\beta}}_0$, $\hat{\boldsymbol{\beta}}$, and $\tilde{\boldsymbol{\beta}}$, respectively where the robust estimates can be either M-estimates or MM-estimates. Denote the information matrix by $\hat{\boldsymbol{U}} = \hat{\tau}^{-2}\boldsymbol{Z}^T\boldsymbol{Z}$ and $\hat{\tau}^2$ consistently estimates the asymptotic variance τ^2 of the unconstrained estimator $\hat{\boldsymbol{\beta}}$. Then, the robust Wald test-statistic is given by

$$\begin{aligned}\overline{RW} = \inf_b\{(\hat{\boldsymbol{\beta}} - \boldsymbol{b})^T\hat{\boldsymbol{U}}(\hat{\boldsymbol{\beta}} - \boldsymbol{b}) : \boldsymbol{Rb} = \boldsymbol{0}\} - \\ \inf_b\{(\hat{\boldsymbol{\beta}} - \boldsymbol{b})^T\hat{\boldsymbol{U}}(\hat{\boldsymbol{\beta}} - \boldsymbol{b}) : \boldsymbol{R_1 b} \geq \boldsymbol{0}\}.\end{aligned} \tag{3.12}$$

Let the slope vector $\boldsymbol{\beta}$ be partitioned as $(\boldsymbol{\beta}_{(1)}^T, \boldsymbol{\beta}_{(2)}^T)^T$, where under H_0 $\boldsymbol{\beta}_{(1)}^T$ is unspecified and $\boldsymbol{\beta}_{(2)}^T \in \mathcal{M}$, while under H_1, $\boldsymbol{\beta}_{(1)}^T$ is unspecified and

$\boldsymbol{\beta}_{(2)}^T \in \mathcal{C}$. Then, the robust score test-statistic is given by

$$\begin{aligned}\overline{RS} = \inf_b\{(S(\hat{\boldsymbol{\beta}}_{(2)}) - \boldsymbol{b})^T \hat{\boldsymbol{C}}^{-1}(S(\hat{\boldsymbol{\beta}}_{(2)}) - \boldsymbol{b}) : \boldsymbol{Rb} = \boldsymbol{0}\} - \\ \inf_b\{(S(\hat{\boldsymbol{\beta}}_{(2)}) - \boldsymbol{b})^T \hat{\boldsymbol{C}}^{-1}(S(\hat{\boldsymbol{\beta}}_{(2)}) - \boldsymbol{b}) : \boldsymbol{R_1 b} \geq \boldsymbol{0}\},\end{aligned} \tag{3.13}$$

where $S(\hat{\boldsymbol{\beta}}_{(2)})$ is the corresponding subvector of the scores vector $S(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^n \psi_2(e_i/\hat{\sigma})\boldsymbol{z}_i\boldsymbol{z}_i^T/n$ with $e_i = y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\theta}}$. The information matrix is denoted by $\hat{\boldsymbol{C}} = \{\boldsymbol{M}_{(22.1)}\boldsymbol{V}_{22}\boldsymbol{M}_{(22.1)}^T\}$, where $\hat{\boldsymbol{V}} = \boldsymbol{M}^{-1}\boldsymbol{Q}\boldsymbol{M}^{-T}$ with $\boldsymbol{M} = \sum_{i=1}^n \psi_2'(e_i/\hat{\sigma})\boldsymbol{z}_i\boldsymbol{z}_i^T/n$, $\boldsymbol{Q} = \sum_{i=1}^n \psi_2^2(e_i/\hat{\sigma})\boldsymbol{z}_i\boldsymbol{z}_i^T/n$, $e_i = y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\theta}}$, and $\hat{\sigma}$ is the scale estimate in the unconstrained model.

At this point, it is worth mentioning that the matrices $\boldsymbol{M}$ and $\boldsymbol{Q}$ can be computed both under the null or unconstrained model to get a consistent estimator of the information matrix $\boldsymbol{C}$ in equation 3.13. The choice depends on the main focus of the test. When sequences of local alternatives (approaching the null hypothesis when the sample size grows) are considered it is perfectly valid to estimate $\boldsymbol{M}, \boldsymbol{Q}$ and $\boldsymbol{C}$ under the null model. In this case, the power of these tests is only studied for alternative hypotheses that are close to the null hypothesis. However, despite its computational attractiveness, the power decreases for alternative hypotheses that are further away from the null hypothesis. To avoid this, we shall evaluate the matrices $\boldsymbol{M}$ and $\boldsymbol{Q}$ under the unconstrained model.

The LRT statistic is defined by

$$\overline{RF} = \hat{\lambda}^{-1}[\inf_b\{L(\boldsymbol{b}, \hat{\sigma}) : \boldsymbol{Rb} = \boldsymbol{0}\} - \inf_b\{L(\boldsymbol{b}, \hat{\sigma}) : \boldsymbol{R_1 b} \geq \boldsymbol{0}\}], \tag{3.14}$$

where $L(\boldsymbol{b}, \hat{\sigma}) = \sum_i^n \rho_2(e_i/\hat{\sigma})$ with $e_i = y_i - \hat{\alpha} - \boldsymbol{x}_i^T\boldsymbol{b}$ is the loss function in M- and MM-estimation and let $\hat{\lambda}$ be the asymptotic covariance matrix standardizing constant which equals $\hat{\lambda} = 2^{-1}(n-p)^{-1}\{\Sigma\psi_2^2(e_i/\hat{\sigma})\}$ $\{n^{-1}\Sigma\psi_2'(e_i/\hat{\sigma})\}^{-1}$.

3.3.3 How to find the null-distribution

The null distribution of each of these test-statistics takes the form of a mixture of χ^2-distributions. In particular, the asymptotic null distribution of the test-statistics is given by

$$\Pr(T \geq t \mid \boldsymbol{R\theta} = \boldsymbol{0}) \simeq \sum_{i=0}^{q} w_i(q, \boldsymbol{\Sigma})\ \Pr(\chi^2_{(r-q+i)} \geq t), \tag{3.15}$$

where T is any of the test-statistics given in Equations 3.9 to 3.14 and $\mathbf{\Sigma}$ equals the covariance matrix. It is important to note that the calculation of the mixing weights w_i is invariant for positive constants such like τ^2 (known or unknown) (Silvapulle & Sen, 2005, p. 32).

Closed form expressions for the mixing weights $w_i(q, \mathbf{\Sigma})$ can be found in Gouriéroux, Holly, and Monfort (1982); Kudô (1963) and Shapiro (1988) for $q \leq 4$. The exact computation of the weights for $q > 4$ is a difficult task in general because the weights can no longer be expressed in closed form. An exception is when the block in the information matrix associated with the inequality constraint parameters is diagonal. In this case the weights follow a Binomial distribution with q trials (i.e., the number of inequality constraints) and probability of success equal to 0.5 (Gouriéroux et al., 1982). For the case of correlated parameter estimates, the weights can be approximated by using the multivariate normal probability distribution function with additional Monte Carlo steps (Grömping, 2010) or they can be computed easily and sufficiently precise by Monte Carlo simulation (Silvapulle & Sen, 2005; Wolak, 1989). Note that the p-value can also be computed directly using parametric or non-parametric bootstrap (Silvapulle & Sen, 2005).

3.4 Simulation study

3.4.1 Design of the simulation study

We generated 2999 samples of sizes $N = 30, 50, 75, 100, 200, 400$ according to the linear regression model $\mathbf{y} = \mathbf{1} + \mathbf{Z}\boldsymbol{\beta} + \mathbf{e}$, where $\boldsymbol{\beta} \in \mathbb{R}^4$, with $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_4)$, independent from $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, 1)$. The vector with regression parameters was set to $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4) = (0, d, d, d)$, where d was varied to obtain samples from the null hypothesis ($d = 0$) and from alternative hypotheses ($d = 0.05, 0.10, \ldots, 0.5$). We restricted the regression coefficients to be unconstrained, positively-constrained ($\beta_i \geq 0, i = 1, 2, 3, 4$) or order-constrained ($0 \leq \beta_1 \leq \beta_2 \leq \beta_3 \leq \beta_4$) and estimated the coefficients via OLS-, M- and MM-estimation. To investigate the performance of the different hypothesis tests we considered three hypotheses, namely the unconstrained hypothesis ($\mathrm{H}_{\mathrm{unc}}$), the positively-constrained hypothesis $\mathrm{H}_{\mathrm{pos}} : \beta_i \geq 0, i = 2, 3, 4$ and the order-constrained hypothesis $\mathrm{H}_{\mathrm{order}} : 0 \leq \beta_2 \leq \beta_3 \leq \beta_4$.

To examine the robustness of the estimators, we investigated the impact of leverage points on the performance of the estimators. To generate contaminated samples, ten percent of the values of the first column of $\boldsymbol{Z}$ were replaced by observations following a $\mathcal{N}(5, 0.1^2)$ distribution while the corresponding response $\boldsymbol{y}$ had a $\mathcal{N}(\eta, 0.1^2)$ distribution, with η taking the values $0, 1, 2, 3 \ldots, 16$, respectively. Larger values of η yield more severe bad leverage points. Note that the contaminated variable is not involved in the hypotheses. This ensures that the contamination affects both the estimation of the parameters under the null- and alternative hypothesis. A similar simulation design is used in Salibián-Barrera, Van Aelst, and Yohai (2014). All results are obtained using the `R` package `restriktor` (version 0.1-80). The `R` code to run the simulations is given in Appendix H.

3.4.2 (Root) mean squared errors

To investigate the robustness of the estimators, Figure 3.1 shows the influence of the outlier configurations on the root mean squared error (RMSE) of the estimators for β_1 and the sample-size $N = 50$. From the plot we see that the MM-estimator (dashed-lines) is the more 'precise' estimator. Its good performance for large values of η is due to the redescending weight-function that is used, which ensures that large residuals get weight zero. On the other hand, the OLS-estimator (solid lines) is clearly non-robust. Its RMSE continues to increase with η. This is also the case for the M-estimator (dotted-lines), except when the regression coefficients are subject to order constraints. For all three estimators, incorporating the order-constraints yields a substantial improvement of the RMSE compared to the corresponding positively-constrained and unconstrained estimators.

To further compare the robustness of the estimators, Table 3.3 gives for all regression parameters the relative mean squared errors (MSE) for uncontaminated samples and contaminated samples for different values of η, $d = 0$ and sample sizes $N = 50$ and $N = 100$. A relative MSE less than 1 indicates that the robust estimator is more precise than the OLS-estimator. Moreover, if $\text{MSE}_{\text{MM}} < \text{MSE}_{\text{M}}$ then the MM-estimator is more precise than the M-estimator. The results show that the OLS-estimator always performs best for uncontaminated samples. However,

with contaminated samples its performance quickly deteriorates when $\eta \geq 6$. In the presence of bad leverage points, MM-estimation outperforms both OLS and M-estimation. Again, order-constraints outperforms the positively-constrained and unconstrained estimators.

3.4.3 Size and (adjusted) power

The exact finite sample distributions of the non-robust F, LR and score test-statistics based on OLS-estimates and normally distributed errors, are a mixture of F distributions under the null hypothesis (Wolak, 1987). In agreement with Silvapulle (1992b), we found that these mixtures of F distributions also better approximate the tail probabilities of the robust tests than their asymptotic distributions. Therefore, the size and power values presented in this section are based on mixtures of F distributions.

First, we investigated the size of the robust and non-robust tests. Figures 3.2a, 3.2b and 3.2c show the influence of the sample-size on the size of the tests when the samples do not contain any irregularities and with a nominal size of 5%. The results show that the accuracy of the tests increases with the sample-size and that the empirical sizes are close to the nominal size for sufficiently large samples. The robust and non-robust F test-statistics are the most accurate tests, even in small samples. On the other hand, the robust Wald and robust score tests are too liberal in smaller samples. For all test statistics the improvement is substantial for constrained hypotheses, where the order-constrained hypothesis again outperforms the positively-constrained hypothesis.

To investigate the power we varied the value of the parameter d. To make the power values comparable, we computed size-adjusted power levels. This adjustment ensures that the empirical level is 0.05 for all tests. The results for $N = 50, 100$, and 200 are shown in Figures 3.2d to 3.2l. As expected, the OLS-tests have the highest power in this setting. The difference is largest for $N = 50$. The robust F-test is the best performing robust test even in small samples. In the unconstrained setting, the robust tests perform somewhat worse than the OLS-tests but these differences become smaller in the constrained settings. Note that the improvement of the order-constrained results are perhaps less severe than expected but this is because we sampled from a model where β_2 to β_4 are taken to be equal.

To investigate the robustness of the tests, Figure 3.3a to 3.3l show the influence of the outlier configurations on both the size ($d = 0$) and power of the unconstrained, the positively-constrained and the order-constrained hypothesis tests for the case $d = 0.10, 0.20, 0.30$ and $N = 100$. The results show that in these settings the size for OLS-tests is hardly affected by the contamination. The size for M- and MM-tests are somewhat more liberal but the results are not alarming, except for the most damaging outlier configurations $\eta = 6$ and $\eta = 7$ for the robust Wald statistic based on MM-estimation. For all tests the improvement is substantial for constrained hypotheses, where the order-constrained hypothesis again outperforms the positively-constrained hypothesis.

Figures 3.3d to 3.3l show that only MM-tests are capable of maintaining high power, while the power for the M- and OLS-tests drops severely with increasing value of η. Again, the improvement is substantial for constrained hypotheses, where the order-constrained hypothesis outperforms the positively-constrained hypothesis.

Overall, we can conclude that the MM-based F-test performs good in terms of size and power in the presence and absence of contamination, except for the situation of the most damaging outlier configuration ($\eta = 7$). In this case the MM-based score test performs best. The OLS- and M-tests are size robust but their power free-falls towards zero for extreme outlier configurations. However, robustness does not come for free but at the expense of a larger sample-size to maintain equal size and power.

3.5 Illustrative example

In the aftermath of a burn event and subsequent hospitalization, both children and parents may experience traumatic stress reactions. Pediatric burn research focuses on this impact of a burn event on parents and between parents and their child (Bakker, Van der Heijden, Van Son, & Van Loey, 2013). Several predictors have been found to be related to parental post-traumatic stress symptoms (PTSS), such as parental emotions in relation to the burn event (e.g., guilt) and the percentage of total body surface area burned (TBSA) (De Young, Hendrikz, Kenardy, Cobham, & Kimble, 2014; Hall et al., 2006).

The data in our example are based on two cohort studies in children

from 0 to 4 and 8 to 18 years old with burns and their parents (e.g., Bakker et al., 2013 and Egberts et al., 2016). For illustrative reasons we focus only on the data provided by the mother. The final sample consists of mothers of 278 children. The response variable is parental post-traumatic stress symptoms (PTSS) and was measured with the Impact of Event Scale. Moreover, for the current illustration we included five predictor variables in the dataset: a child's gender (0 = boys, 1 = girls) and age, the estimated percentage total body surface area affected by second or third degree burns (i.e., TBSA, with a range of 1–72% in the current sample), and the parent's guilt [0–4] and anger [0–4] feelings in relation to the burn event.

Clinical evidence and a previous study suggest that mothers may report higher PTSS levels for girls compared to boys (McGarry et al., 2013). Hence, we are interested in whether the gender-effect increases for simultaneously higher levels of guilt, anger and TBSA. In other words, we are interested in the covariates-conditional effects (Mayer, Dietzfelbinger, Rossel, & Steyer, 2016) of gender on PTSS. A prominent approach to estimate conditional effects is based on multiple regression with interactions. The model with interactions can be written as a linear function

$$\begin{aligned}\text{PTSS} \sim \alpha + \beta_1\text{gender} + \beta_2\text{age} + \beta_3\text{guilt} &+ \beta_4\text{anger} + \beta_5\text{TBSA} \\ &+ \beta_6\text{gender} \times \text{guilt} \\ &+ \beta_7\text{gender} \times \text{anger} \\ &+ \beta_8\text{gender} \times \text{TBSA}.\end{aligned}$$

The conditional effects can be obtained at certain values of the covariates. We selected three different values for the covariates guilt, anger and TBSA, namely a small, a medium and a large level. For a small level, we chose the values 0, 0, 1 for guilt, anger and TBSA respectively. For a medium level we chose their mean values which are 1.525, 1.309, and 8.354, respectively, and for a large level we chose 4, 4, and 35, respectively. Note that these values are chosen for illustrative reasons. Different chosen values may result in a different conclusion.

In contrast to common hypothesis tests which are usually about model parameters (i.e., regression coefficients), effects are defined as a function of the model parameters. The resulting three effects can be calculated as

follows and each effect reflects a mean difference between boys and girls.

$$\begin{aligned} \text{Effect1} &= \beta_1 + \beta_6 0 + \beta_7 0 + \beta_8 1 \\ \text{Effect2} &= \beta_1 + \beta_6 1.525 + \beta_7 1.309 + \beta_8 8.354 \\ \text{Effect3} &= \beta_1 + \beta_6 4 + \beta_7 4 + \beta_8 35. \end{aligned} \tag{3.16}$$

Since, we expect that the gender differences would increase for simultaneously higher levels of guilt, anger and TBSA, the hypothesis of interest is defined as

$$H_1 : \text{Effect1} \leq \text{Effect2} \leq \text{Effect3}. \tag{3.17}$$

The matching constraint matrix $\boldsymbol{R_1}$ can be written as

$$\boldsymbol{R_1} = \begin{bmatrix} \alpha & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 & \beta_6 & \beta_7 & \beta_8 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1.525 & 1.309 & 7.354 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2.475 & 2.691 & 26.646 \end{bmatrix}, \tag{3.18}$$

where the first row refers to the constraint Effect1 $\leq$ Effect2 and the second row to Effect2 $\leq$ Effect3. Manually constructing the constraints matrix as shown in Equation 3.18 can be a complex task. Fortunately, the `R` package `restriktor` can be used for constrained estimation and inference for linear models and allows for easy specification of the constraints. The `R`-code with the model-syntax and constraint-syntax can be found in Appendix I.

Based on outlier diagnostics we identified 12 irregular observations in the data (approximately 4.7% of the data). The diagnostic results are displayed in Figure 3.4. The figure reveals 12 (high)-leverage points which were identified with robust Mahalanobis distances larger than the 99.5% quantile of a χ^2_8 distribution. Therefore, robust estimation of the linear model would be a natural choice. Otherwise we may draw misleading conclusions.

Table 3.1: Constrained effect estimates

Estimator	Effect1		Effect2		Effect3
OLS	3.448	$\leq$	3.458	$\leq$	8.002
MM	3.590	$\not\leq$	3.590	$\leq$	7.158
M	3.589	$\not\leq$	3.589	$\leq$	7.161

In Table 3.1 the effect-estimates for the different regression methods are presented. The key difference is that in case of OLS-estimation the constraints are in line with the data, while in case of M- and MM-estimation the first constraint (Effect1 $\leq$ Effect2) is active. The latter means that the value for Effect 1 is fixed on the boundary value that is still in agreement with the constraint. The practical implication of the results is that in case of OLS-estimation, evidence is found in favor of the order-constrained hypothesis, while for M- and MM-estimation one imposed constraint is not supported by the data. A test is needed to determine whether the violation is severe enough to not reject the null-hypothesis.

To test the order-constrained hypothesis, we used H_0 : Effect1 = Effect2 = Effect3 as competing null-hypothesis. The test-statistics are computed as discussed in Equations 3.9 to 3.14. To obtain the p-values, the weights in the mixtures were calculated by using the multivariate normal distribution function with additional Monte Carlo steps. The results are summarized in Table 3.2. First, the power gain for the constrained tests is clearly visible as the p-values for the unconstrained tests are all bigger than their constrained counterpart. Second, all IC non-robust results are significant, while all IC robust results are not significant. Given that the data are clearly contaminated, as shown in Figure 3.4, we proclaim that the robust results are more reliable. The example illustrates that we may draw misleading conclusions if we ignore the presence of contamination. In particular, we would be led to believe that the data do provide enough evidence that the gender-effect increases for higher levels of guilt, anger and TBSA, while this is *not* supported by careful analysis of the data.

3.6 Summary and discussion

We investigated the performance of inequality constrained (IC) OLS-, M- and MM-estimators when the data are contaminated. The mean squared error (MSE) indicates that MM-estimation produces the most precise estimates. On the other hand, the MSE for the OLS- and the M-estimator can be seriously affected by contamination in the data and increase rapidly for higher levels of contamination (bad leverage points). For all estima-

Table 3.2: Results from the illustration.

		p value	
	Test-statistic	constrained	unconstrained
OLS-estimation:			
$\overline{F}$	5.445	0.040	0.068
$\overline{LR}$	5.571	0.038	0.062
$\overline{S}$	5.377	0.042	0.068
M-estimation:			
$\overline{RF}$	3.289	0.123	0.194
$\overline{RW}$	3.244	0.125	0.194
$\overline{RS}$	4.488	0.065	0.105
MM-estimation:			
$\overline{RF}$	3.295	0.122	0.193
$\overline{RW}$	3.253	0.125	0.195
$\overline{RS}$	4.494	0.065	0.105

tors, it holds that the MSE improves most if the regression coefficients are subject to order constraints.

We mainly investigated the performance of IC (non)-robust likelihood ratio type (LRT), Wald/F, and score tests in terms of size and power. We found that all non-robust tests are size accurate and yield the highest power for uncontaminated samples, as could be expected. The robust LRT test based on M- and MM-estimates is the most accurate robust tests, except when $\eta =$ the most damaging outlier configuration. In this situation, the robust score test performs best. While the non-robust tests are size robust in our contamination settings, their power is severely affected by contaminated ($\eta > 7$) samples. This is also the case for the M-tests. Only MM-tests are capable of maintaining high power, where the robust LRT-test performs adequately. Again, in case of the most damaging outlier configuration, the score test performs best. In addition, order/inequality constraints have a positive effect in diminishing (extreme) outliers in the analyses, where order constraints outperform positive constraints. To improve the size in smaller samples, the residual bootstrap might be a good

alternative estimator. However, the residual bootstrap is not very robust in the presence of outliers. In future research, it should be investigated how alternative methods such as the fast and robust bootstrap (Salibián-Barrera, 2005) can be adapted to the case with constrained hypotheses.

We used an empirical example about child and parental adjustment following a pediatric burn event to show that ignoring outliers may result in spurious conclusions regarding the direction of the effects. We like to emphasize that the application of constrained statistical inference is not limited to the context of burns data. For example, Rosen and Davidov (2012) discusses the constrained linear mixed model applied to the natural history of hearing loss, and Van de Schoot and Strohmeier (2011) discusses the constrained structural equation model applied to psychosocial data.

In the literature, two types of hypothesis tests are often described for testing IC hypotheses, which are often denoted as hypothesis test Type A and hypothesis test Type B (see, e.g. Silvapulle & Sen, 2005). We focused only on hypothesis test Type A where the null hypothesis contains equality constraints and the alternative hypothesis contains inequality constraints. As mentioned earlier, rejecting the null-hypothesis does not mean that the constrained hypothesis is true. Therefore, in practice we often evaluate hypothesis test Type B as well. In hypothesis test Type B, the null hypothesis contains inequality constraints and is tested against the unconstrained hypothesis (some constraints may be preserved in the alternative hypothesis). However, evaluating the power of hypothesis test Type B is less straightforward then evaluating the power of hypothesis test Type A. There is not one obvious choice for the population parameters, since samples are drawn from the unconstrained model and any selected parameters will be arbitrary to some extent. Notwithstanding this, hypothesis test Type B plays a primarily role in constraint misspecification. Results from a previous simulation study (Vanbrabant et al., 2015) have shown that small deviations have only a minor impact on the power. In practice we recommend to evaluate hypothesis test Type B as well to catch severe constraint violations.

In conclusion, many researchers have a-priori knowledge about the order of the parameters in their statistical model. Including order/inequality constraints in the hypothesis has major benefits, such as testing the hypothesis of interest more directly and a substantial gain in power. Nevertheless, ignoring outliers and/or high-leverage points in the analysis may

result in severe power loss and biased estimates. Therefore, in the presence of constraints and data contamination, we advise to use IC robust techniques. Moreover, these methods are now available in the user-friendly `R` package **restriktor** and ready to be used.

Algorithm 1 Iteratively Reweighted Constrained Least Squares optimization steps.

for (iter in 1L:maxit) **do**
    w ← psi.bisquare(resid / scale) ▷ compute Tukey's bisquare weights
    W ← diag(sqrt(w)) ▷ matrix with the weights on the diagonal
    Dmat ← t(X) W X ▷ matrix to be minimized
    dvec ← t(X) W y ▷ vector to be minimized
    β_{restr} ← SOLVE.QP(Dmat, dvec, Amat, bvec, meq) ▷ call quadratic optimizer
    resid.new ← $y - X\beta_{restr}$ ▷ compute residuals under the constraints
    if (abs(resid - resid.new) ≤ absval) **then** ▷ check for convergence
        break
    else
        resid ← resid.new
    end if
end for

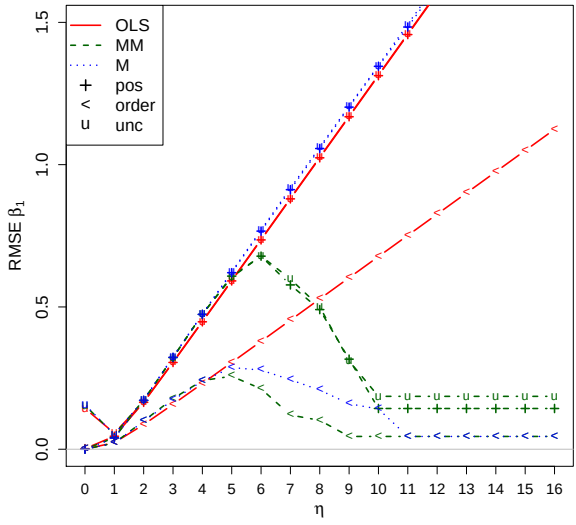
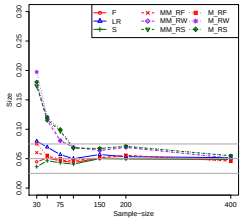


Figure 3.1: The influence of different outlier configurations η on the RMSE for β_1, for $N = 50$.

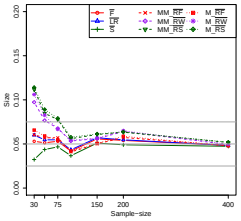
Table 3.3: Relative MSE for $N = 50$ and $N = 100$, and $d = 0$.

			$N = 50$					$N = 100$				
	$\geq\leq$	Relative MSE	β_0	β_1	β_2	β_3	β_4	β_0	β_1	β_2	β_3	β_4
Uncontaminated	unc	MSE_M / MSE_OLS	0.999	1.021	1.049	1.034	1.006	1.006	1.059	1.068	1.035	1.028
		MSE_MM / MSE_OLS	0.998	1.019	1.039	1.035	1.002	1.006	1.058	1.064	1.035	1.031
	pos	MSE_M / MSE_OLS	0.997	1.066	1.036	1.037	1.030	1.005	1.042	1.055	1.027	0.993
		MSE_MM / MSE_OLS	0.996	1.050	1.023	1.022	1.023	1.005	1.034	1.050	1.030	0.996
	order	MSE_M / MSE_OLS	0.997	1.037	1.069	1.062	1.031	1.005	0.989	1.020	1.034	1.007
		MSE_MM / MSE_OLS	0.995	1.039	1.050	1.062	1.026	1.006	0.999	1.031	1.036	1.010
$\eta = 3$	unc	MSE_M / MSE_OLS	1.003	1.065	1.049	1.017	0.997	1.011	1.067	1.032	1.020	0.989
		MSE_MM / MSE_OLS	1.002	1.057	1.042	1.023	0.991	1.010	1.065	1.027	1.015	0.997
	pos	MSE_M / MSE_OLS	1.002	1.074	1.072	0.982	1.013	1.009	1.070	0.988	0.994	0.974
		MSE_MM / MSE_OLS	1.002	1.069	1.036	0.974	1.011	1.008	1.066	0.983	0.993	0.980
	order	MSE_M / MSE_OLS	1.008	1.083	1.083	1.083	1.077	1.011	1.059	1.059	1.059	1.055
		MSE_MM / MSE_OLS	1.010	1.117	1.117	1.116	1.107	1.017	1.093	1.093	1.093	1.088
$\eta = 7$	unc	MSE_M / MSE_OLS	1.005	1.035	1.027	1.045	1.018	1.010	1.034	1.014	1.058	1.016
		MSE_MM / MSE_OLS	0.902	0.703	0.889	0.980	0.931	0.900	0.605	0.837	0.907	0.958
	pos	MSE_M / MSE_OLS	1.005	1.037	1.025	1.014	1.037	1.009	1.034	1.015	1.058	1.016
		MSE_MM / MSE_OLS	0.907	0.680	0.952	0.831	0.883	0.899	0.601	0.765	0.882	1.009
	order	MSE_M / MSE_OLS	0.775	0.499	0.501	0.507	0.564	0.788	0.512	0.514	0.552	0.534
		MSE_MM / MSE_OLS	0.719	0.176	0.181	0.198	0.323	0.726	0.156	0.167	0.190	0.256
$\eta = 9$	unc	MSE_M / MSE_OLS	1.007	1.027	1.038	1.056	1.019	1.008	1.024	1.019	1.069	1.024
		MSE_MM / MSE_OLS	0.812	0.282	0.662	0.739	0.790	0.812	0.151	0.614	0.710	0.786
	pos	MSE_M / MSE_OLS	1.007	1.028	1.038	1.034	1.035	1.008	1.024	1.023	1.072	1.015
		MSE_MM / MSE_OLS	0.812	0.235	0.628	0.606	0.720	0.809	0.072	0.568	0.654	0.865
	order	MSE_M / MSE_OLS	0.658	0.211	0.214	0.223	0.293	0.667	0.146	0.152	0.167	0.210
		MSE_MM / MSE_OLS	0.649	0.063	0.071	0.095	0.213	0.659	0.051	0.070	0.099	0.162
$\eta = 12$	unc	MSE_M / MSE_OLS	1.007	1.017	1.042	1.067	1.012	1.005	1.016	1.026	1.186	1.034
		MSE_MM / MSE_OLS	0.736	0.109	0.532	0.578	0.581	0.754	0.070	0.488	0.577	0.612
	pos	MSE_M / MSE_OLS	1.006	1.018	1.050	1.059	1.035	1.005	1.015	1.030	1.092	1.016
		MSE_MM / MSE_OLS	0.739	0.076	0.504	0.485	0.525	0.753	0.052	0.444	0.533	0.702
	order	MSE_M / MSE_OLS	0.573	0.045	0.053	0.070	0.156	0.583	0.037	0.051	0.072	0.118
		MSE_MM / MSE_OLS	0.573	0.046	0.051	0.069	0.115	0.583	0.037	0.051	0.072	0.118

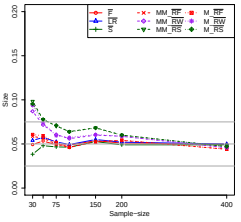
[1] unc = unconstrained, pos = positively-constrained, order = order-constrained.



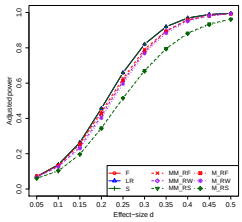
(a) Size, unc



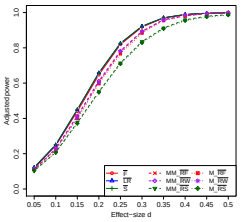
(b) Size, pos



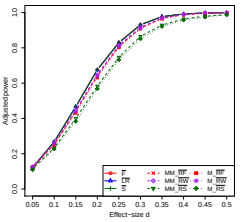
(c) Size, order



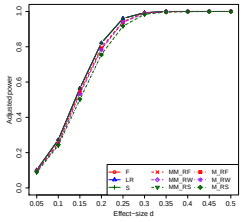
(d) N = 50, unc



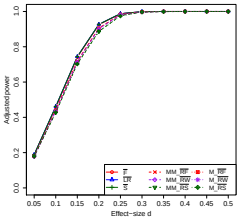
(e) N = 50, pos



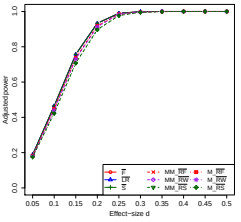
(f) N = 50, order



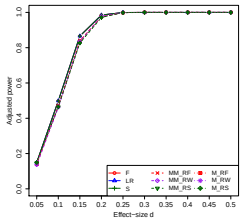
(g) N = 100, unc



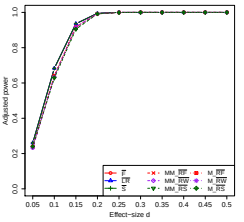
(h) N =100, pos



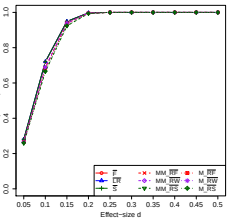
(i) N = 100, order



(j) N = 200, unc

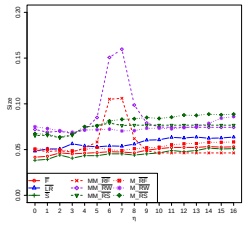


(k) N = 200, pos

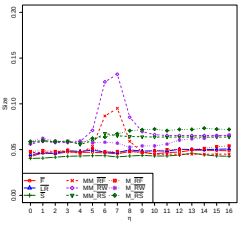


(l) N = 200, order

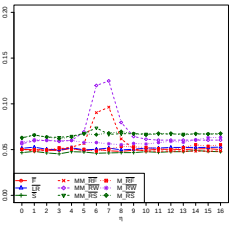
Figure 3.2: The influence of sample size on the size and the influence of effect-size on the adjusted power levels, when no contamination is present.



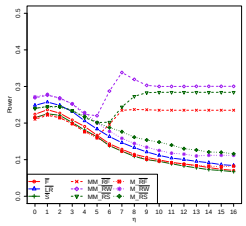
(a) Size, unc



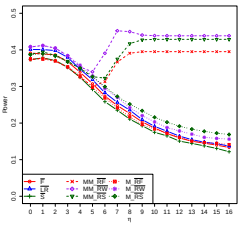
(b) Size, pos



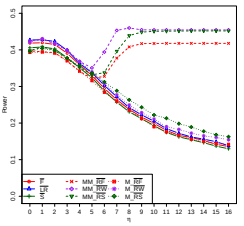
(c) Size, order



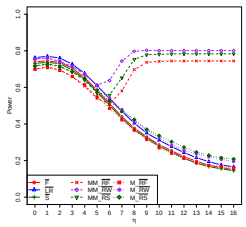
(d) $d = 0.10$, unc



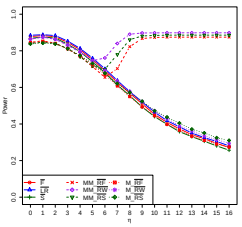
(e) $d = 0.10$, pos



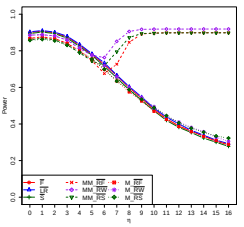
(f) $d = 0.10$, order



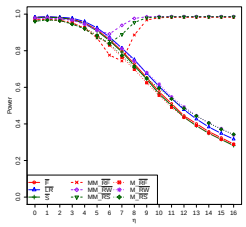
(g) $d = 0.20$, unc



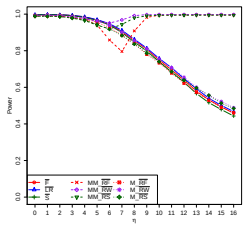
(h) $d = 0.20$, pos



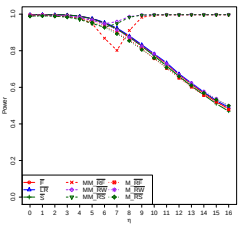
(i) $d = 0.20$, order



(j) $d = 0.30$, unc



(k) $d = 0.30$, pos



(l) $d = 0.30$, order

Figure 3.3: The influence of different outlier configurations η on the size ($d = 0$) and power ($d = 0.10, 0.20, 0.30$), for $N = 100$ and the unconstrained, positively-constrained and order-constrained hypothesis.

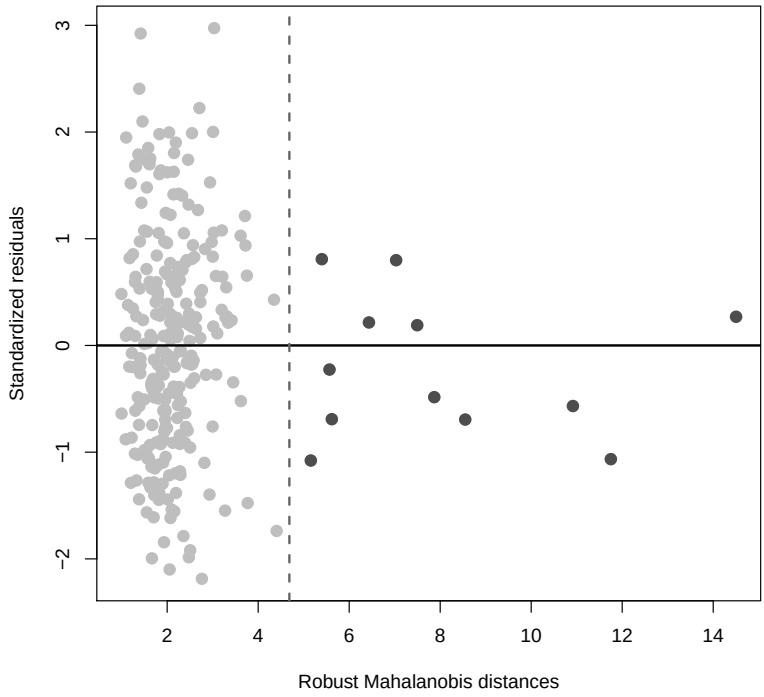


Figure 3.4: Plot of standardized residuals against robust Mahalanobis distances for the burns data. The vertical dashed line indicates the 99.5% quantile of a χ^2_8 distributions. Observations beyond this line are considered as (high)-leverage points.

H

Simulation R-code for N = 50, 10% contamination and order constraints

```
library(restriktor)
library(MASS)

# number of parameters
p <- 4
# 10% contamination
cont <- 0.10
# order constraints
myConstraints <- "x2 > 0; x2 < x3; x3 < x4;"

seed <- 3013073
parallel <- "multicore"
ncpus <- 8

# sample-size
N <- 100
# effect-size
d <- 0
# damaging outlier configurations
```

```
eta <- c(0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16)

result <- list()
nsims <- 2999
pvalues <- matrix(NA, nsims, 9)
for (l in 1:length(eta)) {
  cat("iteration eta =", l, "\n")
  betas <- c(0,d,d,d)

  fn <- function(b) {
    set.seed(seed + b)
    X <- mvrnorm(N, mu = rep(0, p), Sigma = diag(p)
      colnames(X) <- c("x1","x2","x3","x4")
    y <- 1 + X%*%betas + rnorm(N)

    idx <- sample(1:nrow(y), N*cont, replace = FALSE)
    X[idx,1] <- rnorm(length(idx), 5, 0.1)
    y[idx,1] <- rnorm(length(idx), eta, 0.1)
    sim.data <- data.frame(y, X)

    # ols-estimation
    fit.ols <-  lm(y ~ x1 + x2 + x3 + x4, data = sim.data)
    restr1.ols <- iht(fit.ols, constraints = myConstraints, type = "A",
                      test = "F")
    restr2.ols <- iht(fit.ols, constraints = myConstraints, type = "A",
                      test = "LRT")
    restr3.ols <- iht(fit.ols, constraints = myConstraints, type = "A",
                      test = "score")

    # MM-estimation
    fit.mm  <- rlm(y ~ x1 + x2 + x3 + x4, data = sim.data, method = "MM")
    restr1.mm <- iht(fit.mm, constraints = myConstraints, type = "A",
                     test = "F")
    restr2.mm <- iht(fit.mm, constraints = myConstraints, type = "A",
                     test = "Wald")
    restr3.mm <- iht(fit.mm, constraints = myConstraints, type = "A",
                     test = "score")

    # M-estimaion
    fit.m   <- rlm(y ~ x1 + x2 + x3 + x4, data = sim.data, method = "M",
                   psi = psi.bisquare)
    restr1.m <- iht(fit.m, constraints = myConstraints, type = "A",
                    test = "F")
```

```
    restr2.m <- iht(fit.m, constraints = myConstraints, type = "A",
                    test = "Wald")
    restr3.m <- iht(fit.m, constraints = myConstraints, type = "A",
                    test = "score")

    out <- c(restr1.ols$pvalue, restr2.ols$pvalue, restr3.ols$pvalue,
             restr1.mm$pvalue,  restr2.mm$pvalue,  restr3.mm$pvalue,
             restr1.m$pvalue,   restr2.m$pvalue,   restr3.m$pvalue)
    out
  }

  res <- if (ncpus > 1L) {
    parallel::mclapply(seq_len(nsims), fn, mc.cores = ncpus)
  } else {
    lapply(seq_len(nsims), fn)
  }
  error.idx <- integer(0)
  for (b in seq_len(nsims)) {
    if (!is.null(res[[b]])) {
      pvalues[b, 1:ncol(pvalues)] <- res[[b]]
    }
    else {
      error.idx <- c(error.idx, b)
    }
  }
  result[[l]] <- pvalues
}

# compute power
power <- matrix(NA, length(eta), 9)
for (l in 1:length(eta)) {
  for (i in 1:9) {
    power[l,i] <- sum(result[[l]][,i] <= 0.05) / nsims
  }
}
```

I

R-code burns data example

```
library(restriktor)
library(MASS)

## fit unconstrained linear model using OLS-, M- and
## MM-estimation OLS-estimation
fit.ols <- rlm(PTSS ~ gender*guilt + gender*anger +
               gender*TBSA + age, data = burnsData)

# MM-estimation
fit.mm <- rlm(PTSS ~ gender*guilt + gender*anger +
              gender*TBSA + age, data = burnsData,
              method = "MM")

# M-estimation
fit.m <- rlm(PTSS ~ gender*guilt + gender*anger +
             gender*TBSA + age, data = burnsData,
             method = "M", psi = "psi.bisquare")
```

```
# defining the effects and specifying the constraints
constraints <- "Effect1 := gender + 0*gender.guilt +
                                    0*gender.anger +
                                    1*gender.TBSA

                Effect2 := gender + 1.53*gender.guilt +
                                    1.31*gender.anger +
                                    8.35*gender.TBSA

                Effect3 := gender + 4*gender.guilt +
                                    4*gender.anger +
                                   35*gender.TBSA

                Effect1 < Effect2; Effect2 < Effect3"

# compute test-statistic and pvalue.
iht(fit.ols, constraints = constraints, test = "F", type = "A")
iht(fit.mm,  constraints = constraints, test = "F", type = "A")
iht(fit.m,   constraints = constraints, test = "F", type = "A")
```

References

Bakker, A., Van der Heijden, P., Van Son, M., & Van Loey, N. (2013). Course of traumatic stress reactions in couples after a burn event to their young child. *Health Psychology*, *10*(32), 1076–1083. doi: doi:10.1037/a0033983

Barlow, R., Bartholomew, D., Bremner, H., & Brunk, H. (1972). *Statistical inference under order restrictions*. New York: Wiley.

Bartholomew, D. (1961a). Ordered tests in the analysis of variance. *Biometrika*, *48*(3/4), 325–332. doi: doi:10.2307/2332754

Bartholomew, D. (1961b). A test of homogeneity of means under restricted alternatives. *Journal of the royal statistical society. Series B (Methodological)*, *23*(2), 239–281.

De Young, A. C., Hendrikz, J., Kenardy, J. A., Cobham, V. E., & Kimble, R. M. (2014). Prospective evaluation of parent distress following pediatric burns and identification of risk factors for young child and parent posttraumatic stress disorder. *Journal of Child and Adolescent Psychopharmacology*, *1*(24), 9–17. doi: doi:10.1089/cap.2013.0066

Egberts, M. R., van de Schoot, R., Boekelaar, A., Hendrickx, H., Geenen, R., & N.E.E., V. (2016). Child and adolescent internalizing and externalizing problems 12 months postburn: the potential role of preburn functioning, parental posttraumatic stress, and informant bias. *Child & Adolescent Psychiatry*, *25*(7), 791–803. doi: doi:10.1007/s00787-015-0788-z

Gouriéroux, C., Holly, A., & Monfort, A. (1982). Likelihood ratio test, wald test, and kuhn-tucker test in linear models with inequality constraints on the regression parameters. *Econometrica*, *50*, 63–80. doi: doi:10.2307/1912529

Grömping, U. (2010). Inference with linear equality and inequality constraints using R: The package ic.infer. *Journal of statistical software*, *33*, 1–31. doi: doi:10.18637/jss.v033.i10

Hall, E., Saxe, G., Stoddard, F., Kaplow, J., Koenen, K., Chawla, N., ... King, D. (2006). Posttraumatic stress symptoms in parents of children with acute burns. *Journal of Pediatric Psychology*, *31*(4), 403–412. doi: doi:10.1093/jpepsy/jsj016

Hampel, F. (1973). Robust estimation: A condensed partial survey. *Probability theory and related fields*, *27*(2), 87–104.

Hoijtink, H. (2012). *Informative Hypotheses: Theory and Practice for Behavioral and Social Scientists*. Boca Raton, FL: Taylor & Francis.

Huber, P. (1973). Robust regression: asymptotics, conjectures and monte carlo. *The annals of statistics*, *1*(5), 799–821. doi: doi:doi:10.1214/aos/1176342503

Kudô, A. (1963). A multivariate analogue of the one-sided test. *Biometrika*, *50*(3/4), 403–418. doi: doi:10.2307/2333909

Kuiper, R., & Hoijtink, H. (2010). Comparisons of means using exploratory and confirmatory approaches. *Psychological Methods*, *15*(1), 69–86. doi: doi:10.1037/a0018720

Maronna, R., Martin, D., & Yohai, V. (2006). *Robust statistics: Theory and methods*. John Willey and Sons, New York.

Mayer, A., Dietzfelbinger, L., Rossel, Y., & Steyer, R. (2016). The effectliter approach for analyzing average and conditional effects. *Multivariate Behavioral Research*, *51*(2–3). doi: doi:10.1080/00273171.2016.1151334

McGarry, S., Girdler, S., McDonald, A., Valentine, J., Wood, F., & Elliott, C. (2013). Paediatric medical trauma: The impact on parents of burn survivors. *Burns*, *6*(39), 1114–1121. doi: doi:10.1016/j.burns.2013.01.009

Meyer, M., & Wang, J. (2012). Improved power of one-sided tests. *Statistics & probability letters*, *82*(8), 1619–1622. doi: doi:10.1016/j.spl.2012.04.016

Nocedal, J., & Wright, S. (2006). *Numerical Optimization* (2nd ed.; V. Mikosh, S. Resnick, & S. Robinson, Eds.). Spring-Verlag: New York.

Perlman, M. (1969). One-sided testing problems in multivariate analysis. *The Annals of Mathematical Statistics*, *40*(2), 549–567. doi: doi:10.1214/aoms/1177697723

R Development Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. (ISBN 3-900051-07-0)

Robertson, T., Wright, F. T., & Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. New York: Wiley.

Rosen, S., & Davidov, O. (2012). Order-restricted inference for multivariate longitudinal data with applications to the natural history of hearing loss. *Statistics in Medicine*, *31*(16), 1761–1773. doi: doi:10.1002/sim.5335

Rousseeuw, P., & Yohai, V. (1984). Robust regression by means of S-Estimators. In J. Franke, W. Härdle, & D. Martin (Eds.), *Robust and nonlinear time series analysis* (pp. 256–272). Spring-Verlag: New York.

Salibián-Barrera, M., Van Aelst, S., & Yohai, V. (2014). Robust tests for linear regression models based on τ-estimates. *Computational Statistics and Data Analysis*. doi: doi:10.1016/j.csda.2014.09.012v

Salibián-Barrera, M. (2005). Estimating the p-values of robust tests for the linear model. *Journal of statistical planning and inference*, *128*(1), 241–257. doi: doi:10.1016/j.jspi.2003.09.033

Schrader, R., & Hettmansperger, T. (1980). Robust analysis of variance based upon a likelihood ratio criterion. *Biometrika*, *67*(1), 93–101. doi: doi:10.2307/2335321

Shapiro, A. (1988). Towards a unified theory of inequality constrained testing in multivariate analysis. *International Statistical Review*, *56*, 49–62. doi: doi:10.2307/1403361

Silvapulle, M. (1992a). Robust tests of inequality constraints and one-sided hypotheses in the linear model. *Biometrika*, *79*(3), 621–630. doi: doi:10.2307/2336793

Silvapulle, M. (1992b). Robust wald-type tests of one-sided hypotheses in the linear model. *Journal of the American Statistical Association*, *87*(417), 156–161. doi: doi:10.2307/2290464

Silvapulle, M. (1996). Robust bounded influence tests against one-sided hypotheses in general parametric models. *Statistics & probability letters*, *31*(1), 45–50.

Silvapulle, M., & Sen, P. (2005). *Constrained statistical inference: Order, inequality, and shape restrictions*. Hoboken, NJ: Wiley.

Silvapulle, M., & Silvapulle, P. (1995). A score test against one-sided alternatives. *American statistical association*, *90*(429), 342–349. doi: doi:10.2307/2291159

Turlach, B., & Weingessel, A. (2013). quadprog: Functions to solve quadratic programming problems (version 1.5-5). [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=quadprog`

Van de Schoot, R., & Strohmeier, D. (2011). Testing informative hypotheses in SEM increases power: An illustration contrasting classical hypothesis testing with a parametric bootstrap approach. *International Journal of Behavioral Development*, *35*, 180–190. doi: doi:10.1177/0165025410397432

Vanbrabant, L., Van de Schoot, R., & Rosseel, Y. (2015). Constrained statistical inference: sample-size tables for anova and regression. *Frontiers in Psychology*, *5*, 1–8. doi: doi:10.3389/fpsyg.2014.01565

Wolak, F. (1987). An exact test for multiple inequality and equality constraints in the linear regression model. *Journal*

of the American statistical association, *82*(399), 782–793. doi: doi:10.1080/01621459.1987.10478499

Wolak, F. (1989). Testing inequality constraints in linear econometric models. *Journal of Econometrics*, *41*(2), 205–235. doi: doi:10.1016/0304-4076(89)90094-8

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

4

Evaluating an order-constrained hypothesis against its complement using the GORIC

An order-restricted information criterion such as the GORIC can be used to rank the competing order-restricted hypotheses from best to worst. The unconstrained hypothesis, where no restrictions are placed on the model parameters is usually included as safeguard in the set of hypothesis to avoid selecting a weakly supported hypothesis. The GORIC values themselves are not interpretable. To improve the interpretation regarding the strength, GORIC weights and related evidence ratios can be computed. However, if the unconstrained hypothesis is used as competing hypothesis, the evidence ratio is not affected by sample-size or effect-size in case the hypothesis of interest is (also) in agreement with the data. In practice, this means that strong support for the order-constrained hypothesis is not reflected by a high evidence ratio. Therefore, we introduce the evaluation of an order-constrained hypothesis against its complement using the GORIC (weights). In a small simulation study, we show that

the evidence ratio for the order-constrained hypothesis versus the complement increases for larger samples and effect-sizes, while the evidence ratio for the order-constrained hypothesis versus the unconstrained hypothesis remains bounded. An empirical example about facial burn injury illustrates our method and shows that using the complement as competing hypothesis results in much more support for the hypothesis of interest than using the unconstrained hypothesis as competing hypothesis.

4.1 Introduction

Consider the hypothesis $H_1 : \mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4$, where μ reflects the population mean for each group. This form of hypothesis is known as an order-constrained hypothesis or informative hypothesis (Hoijtink, 2012) because the order of the means is restricted based on theory and/or academic reasoning. To evaluate such order-constrained hypothesis, three methods can be distinguished, i.e. hypothesis testing, model selection using information criteria and Bayesian model selection. In this current article, we focus on model selection using information criteria. The AIC (Akaike, 1998) is probably the most familiar and widely used information criterion employed in the social and behavioral sciences. Nevertheless, the AIC is not suitable when the model parameters (e.g., means and regression coefficients) are subject to order constraints. A modification of the AIC that can deal with simple order constraints in the exponential family was proposed by Anraku (1999) and is called the order-restricted information criterion (ORIC). Kuiper, Hoijtink, and Silvapulle (2011) generalized the ORIC (GORIC) to accommodate any linear inequality constraints in multivariate normal linear models (except for range restrictions, which bounds a parameter to a specific interval, e.g., $-1 \leq \mu \leq 1$). Information criteria like the AIC, ORIC and GORIC are calculated as minus two times the log-likelihood plus twice a penalty term value. The difference between the methods is in calculating the penalty term value, which is less straightforward to compute in case of order constraints.

The evaluation of an order-constrained hypothesis (e.g., H_1) requires a competing hypothesis. To avoid selecting a weakly supported hypothesis as the best one, the unconstrained hypothesis H_u is usually included as a safeguard in the set of M hypotheses. Sometimes researchers have another

hypothesis of interest, for example $H_2 : \mu_1 \leq \mu_2 \leq \mu_3 = \mu_4$ but often they do not have such a specific competing hypothesis. In that case, only H_u is used as competing hypothesis. Therefore, we focus solely on the set of hypotheses with one order-constrained hypothesis H_m and H_u. The hypothesis with the lowest GORIC value is the preferred one. The GORIC values themselves are not interpretable and only the differences between the values can be inspected. To improve the interpretation, so-called GORIC weights (w_m) can be computed, which are comparable to the Akaike weights (Burnham & Anderson, 2002). The GORIC weight w_m represents the relative likelihood of hypothesis m given the data and the set of M hypotheses (Kuiper, 2011, p. 106). For example, if we compare hypothesis H_1 against hypothesis H_u, we can examine the ratio of the two corresponding weights, that is w_1/w_u. This relative evidence reflects how many times hypothesis H_1 is more likely than hypothesis H_u.

However, if the order-constrained hypothesis of interest H_m is in agreement with the data, increasing the sample-size and/or effect-size does not affect the relative evidence if the unconstrained hypothesis is used as competing hypothesis. In that case, both hypotheses H_m and H_u are in line with the data, since H_u is always in line with the data, and consequently both hypotheses have the same maximized likelihood value. Then, the difference in GORIC values equals the difference in penalty term values, which are independent of sample-size and effect-size. The latter case is illustrated in Figure 4.1, where we generated 500 data sets according to an ANOVA model with two uncorrelated ordered means $H_3 : \mu_1 \geq \mu_2$ with a sample-size of $n = 50$ per group and various effect-sizes f. The effect-size f is defined according to Cohen (1988, pp. 274–275). The results show that at first the mean evidence ratio of w_3/w_u increases for increasing effect-sizes and that afterwards it stabilizes at an upper-bound value of approximately 1.65. It is at this point that the data are in agreement with H_3 and thus the maximized log-likelihood values of H_3 and H_u are the same. The boundary value equates the exponential difference of the penalty term values between H_3 and H_u, that is, $\exp(2.00 - 1.50) = \exp(0.50) \approx 1.65$; as will become clear later on. Consequently, strong support for the order-constrained means is not expressed in a high relative evidence and many research questions may be erroneously dismissed as irrelevant.

Therefore, the objective of this study is to show that this upper bound

issue can be solved by replacing the unconstrained hypothesis by the complement of the hypothesis of interest. The complement is defined as $H_c = \neg H_m$, where $\neg$ denotes 'not'. For example, for the order-constrained hypothesis H_1 with 4 means there are 24 ways (i.e., $4! = 4 \times 3 \times 2 \times 1$) in which the four means can be ordered. Hypothesis H_1 consists of 1 of these 24 combinations, therefore the complement represents the 24 - 1 = 23 remaining ways in which the four means can be ordered [1]. In a small simulation study, we show for larger sample-sizes and effect-sizes that, averaged over the samples, the relative evidence for H_m versus H_c (i.e., w_m/w_c) is boundless and thus the evidence for a true hypothesis increases with increasing sample-size and effect-size. An empirical example about facial burn injury illustrates the application of this method.

The remainder of this article is organized as follows. First, we provide some technical background about how to compute the GORIC and the corresponding penalty term value for the unconstrained hypothesis and the order-constrained hypothesis. Next, we show how to evaluate an order-constrained hypothesis against its complement using the GORIC (weights). Third, we investigate the performance of the relative evidence weight w_m/w_c by means of a simulation study. Fourth, we illustrate our method with an empirical example. Finally, we give some concluding remarks and recommendations.

4.2 Technical background

The results given in this part are for the linear regression model, where the regression coefficients are subject to linear inequality and/or linear equality constraints.

4.2.1 Linear model and order-constrained hypotheses

Consider the standard linear regression model,

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\theta} + \epsilon_i, \ i = 1, \ldots, n, \tag{4.1}$$

[1] Note that it is often a cumbersome or even impossible task to write up all possible combinations that belong to the complement, since the number of combinations increases excessively with the number of parameters.

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$ is the parameter vector of interest,
$\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ are vectors of predictor variables [2], and
$\epsilon_i = (\epsilon_1, \ldots, \epsilon_n)^T$ are normally distributed random errors, that is $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Let the unconstrained maximum likelihood estimates (mle's) denoted by $\hat{\boldsymbol{\theta}}$ and the order-restricted mle's denoted by $\tilde{\boldsymbol{\theta}}$.

We consider three types of hypotheses, namely $H_u : \boldsymbol{\theta} \in \mathbb{R}^p$, where $\mathbb{R}^p$ is the p-dimensional Euclidean space, $H_m : \boldsymbol{\theta} \in \mathcal{C}$, where $\mathcal{C}$ is also a space in $\mathbb{R}^p$ and is a (reallocated) closed convex cone, and $H_c : \neg H_m$, which is not necessarily a (reallocated) closed convex cone. Since, most applications only involve linear constraints, we only consider linear hypotheses, then H_m can be written in the more familiar form $\boldsymbol{R\theta} \geq \mathbf{r}$ and H_c can be written as $\neg\boldsymbol{R\theta} \geq \mathbf{r}$ (which is mostly not solely equal to $\boldsymbol{R\theta} \leq \mathbf{r}$). Let $\boldsymbol{R}$ be a matrix of order $q \times p$ with known constants and of rank($\boldsymbol{R}$) $= q$, where $\boldsymbol{R}$ is of full row-rank if $q \leq p$, and $\mathbf{r}$ an $q \times 1$ vector with known constants (often this vector contains zeros). Let us assume that the q restrictions are $q_1 \geq 0$ inequality constraints and $q_2 \geq 0$ equality constraints. Then, let $\boldsymbol{R}_1$ be a matrix of order $q_1 \times p$ and $\mathbf{r}_1$ a matrix of order $q_1 \times 1$, and $\boldsymbol{R}_2$ be a matrix of order $q_2 \times p$ and $\mathbf{r}_2$ a matrix of order $q_2 \times 1$, and $\boldsymbol{R} = [\boldsymbol{R}_1^T, \boldsymbol{R}_2^T]^T$ and $\mathbf{r} = [\mathbf{r}_1^T, \mathbf{r}_2^T]^T$.

4.2.2 GORIC

The GORIC for the unconstrained hypothesis H_u is defined as

$$\text{GORIC}_u = -2 \times \text{LL}_u + 2 \times \text{PT}_u, \tag{4.2}$$

where LL_u is the maximized log-likelihood value for the unconstrained hypothesis and the penalty term value is defined as $\text{PT}_u = 1 + p$. Note that GORIC_u equals the AIC for H_u.

The GORIC for the order-constrained hypothesis H_m is defined as

$$\text{GORIC}_m = -2 \times \text{LL}_m + 2 \times \text{PT}_m, \tag{4.3}$$

where LL_m is the maximized log-likelihood value for the order-constrained hypothesis m. The penalty term value equals
$\text{PT}_m = 1 + \sum_{j=0}^{p} LP_j(p, \boldsymbol{\Sigma}, H_m)j$, where $\boldsymbol{\Sigma} = (\mathbf{X}^T\mathbf{X})^{-1}$ is the unscaled

[2] In case of an intercept $x_{i1} = 1$ for all i's and θ_1 is interpreted as the intercept.

covariance matrix [3] of the parameters with $\boldsymbol{X} = (x_1^T, \ldots, x_n^T)^T$ of order n $\times$ p and $LP_j(p, \boldsymbol{\Sigma}, H_m)$ are the level probabilities (chi-bar-square weights) and sum up to one. A level probability LP_j, is the probability that $\tilde{\boldsymbol{\theta}}$ has j levels, which corresponds to j inactive constraints under $\boldsymbol{R\theta} = \mathbf{r}$, where $j = p -$ the number of active constraints. To clarify, in case of an inactive constraint the mle's do not change if the constraint is removed, while the mle's do change if an active constraint is removed. From the above it follows that for q_2 equality constraints (i.e., q_2 constant parameters) and $(p - q_2)$ non-constant parameters, the penalty term value for the unconstrained hypothesis is $\mathrm{PT}_u = 1 + (p - q_2)$, which equals the penalty term value of the AIC. In case of inequality constraints, the exact computation of the level probabilities when $\boldsymbol{\Sigma} \neq \mathbf{I}$ ($\mathbf{I}$ is an identity matrix) and for $q > 4$ is a difficult task in general because the probabilities can no longer be expressed in closed form. Fortunately, the probabilities can be approximated by using the multivariate normal probability distribution function with additional Monte Carlo steps (Grömping, 2010) or they can be computed easily and sufficiently precise by Monte Carlo simulation (Silvapulle & Sen, 2005; Wolak, 1987).

To illuminate the computation of the penalty term value PT_m, consider Figure 4.2a, where the unrestricted parameter space is determined by the two parameters θ_1 and θ_2 and is divided into four quadrants (Q_1 to Q_4). Note that we have only depicted the parameter space between -4 and 4 and not the whole parameter space. If we assume that θ_1 and θ_2 are independent of each other (i.e., $\boldsymbol{\Sigma} = \mathbf{I}$), then each quadrant gets assigned a level probability of 0.25. The permissible gray-shaded area is defined by the order constraints $H_4 : \theta_1 \geq 0, \theta_2 \geq 0$. Then, the probability that $j = 0$, that is, that both constraints are active (i.e., $j = p - 2 = 2 - 2 = 0$) is 0.25 (Q_1). The probability that $j = 1$, that is, that one constraint is active and that the other constraint is inactive (i.e., $j = 2 - 1 = 1$) is $0.25 + 0.25 = 0.50$ (Q_2 and Q_4). The probability that $j = 2$, that is, that both constraints are inactive (i.e., $j = 2 - 0 = 2$) is 0.25 (Q_3). Then, the penalty term value for the order-constrained hypothesis H_4 can be computed as $\mathrm{PT}_4 = 1 + 0.25 \times 0 + 0.50 \times 1 + 0.25 \times 2 = 2$. In addition, consider Figure 4.2b, where the parameter space is restricted by

[3] The calculation of the level probabilities is invariant for positive constants like σ^2 (known or unknown) (Silvapulle & Sen, 2005, p. 32) or even for $\tilde{\sigma}^2$, the order-constrained mle of σ^2.

the order constraint $H_5 : \theta_1 \geq \theta_2$. Since the order constraint divides the unrestricted parameter space into two spaces, Q_1 and Q_2 are now two half-spaces. Again, we assume that $\mathbf{\Sigma} = \mathbf{I}$. To compare with the penalty term for H_4, again we have two parameters (i.e., $p = 2$) but now we only have one order constraint. Since, the minimum number of inactive constraints is equal to $p - q = 2 - 1 = 1$, the probability that $j = 0$, that is, that we have no inactive constraints is 0. This is because, if we impose one order constraint on two parameters, one parameter is allowed to vary freely (i.e., be inactive), while the other parameter is restricted by the value of this free parameter. Stated otherwise, the probability of two active constraints (i.e., $j = 0$) is 0 in case of only one available constraint. The probability that $j = 1$, that is, that the free parameter is inactive and that the order constraint is active is 0.5 (Q_2). The probability that $j = 2$, that is, that the free parameter and the order constraint are inactive is 0.5 (Q_1). Hence, the penalty term value for the order-constrained hypothesis H_5 is computed as $\text{PT}_5 = 1 + 0 \times 0 + 0.5 \times 1 + 0.5 \times 2 = 2.5$.

4.3 The complement

Here we introduce our method for computing the GORIC for the complement of H_m, which is computed as follows

$$\text{GORIC}_c = -2 \times \text{LL}_c + 2 \times \text{PT}_c, \tag{4.4}$$

where LL_c is the maximized log-likelihood value for the complement of H_m and PT_c is the penalty term value. Recall that for the computation of the GORIC value for H_m the constraints are required to be a closed convex cone. However, the complement H_c is in many cases not a closed convex cone. Moreover, it is often not an easy task (or even impossible) to write out the complement. Consequently, the LL_c and the PT_c values cannot be computed directly like we did for the LL_m and the PT_m values. Next, we will show how to compute the LL_c and the PT_c values based on the components determined under H_m and H_u.

To compute the LL_c value, we first need to ascertain whether the constraints in H_m are in line with the data or not. If at least one inequality constraint is violated, then the data are automatically in line with the complement and the LL_c value equals the LL_u value. This is illustrated

for $H_4 : \theta_1 \geq 0,\ \theta_2 \geq 0$ in Figure 4.3a, where the permissible area is Q_1 and the quadrants Q_2, Q_3 and Q_4 form the complement. Since, the unconstrained mle's $\hat{\boldsymbol{\theta}}$ lay in Q_3 (here both constraints are violated), the data are in line with the complement and the LL_c is equal to LL_u. Note that the same applies if the mle's lay in Q_2 or Q_4. On the other hand, if the data are in line with the constraints in H_4, then we have to find the mle's of $\boldsymbol{\theta}$ that are closest to $\boldsymbol{\theta} \in H_c$, given $\boldsymbol{\Sigma}$. Clearly, the solution is on the boundary of the restricted parameter space H_m and is denoted by $\boldsymbol{\theta}_c$. This is illustrated in Figure 4.3b for the bivariate normal distribution and $\boldsymbol{\Sigma} = \mathbf{I}$. This latter is depicted by the round circles of the contour plot, which indicate that the two parameters θ_1 and θ_2 are uncorrelated. As a reminder, the lines of the contour plot correspond to parameter values which have equal log-likelihood values and lines closer to $\hat{\boldsymbol{\theta}}$ result in a higher log-likelihood value, since $\hat{\boldsymbol{\theta}}$ is the value for which the log-likelihood is maximized (without imposing restrictions on the parameters). Since there are many boundary solutions (see thick black lines), we have to search for a solution that has the shortest distance between $\hat{\boldsymbol{\theta}}$ and the two boundaries, given $\boldsymbol{\Sigma}$. Fortunately, we do not have to investigate each point on the thick black lines but only the points $\tilde{\boldsymbol{\theta}}_{c1}$ and $\tilde{\boldsymbol{\theta}}_{c2}$. The point $\tilde{\boldsymbol{\theta}}_{c1}$ is computed by treating the inequality constraint for θ_1 as equality constraint (i.e., $\theta_1 = 0,\ \theta_2 \geq 0$). Analogously, for the point $\tilde{\boldsymbol{\theta}}_{c2}$, where θ_2 is treated as equality constraint (i.e., $\theta_1 \geq 0,\ \theta_2 = 0$). Thus, in total there are q_1 possibilities to be investigated. Notable, in case of equality constraints, all q_2-equalities are 'freed'. The point that results in the highest log-likelihood value, given $\boldsymbol{\Sigma}$, equals the LL_c value (here $\tilde{\boldsymbol{\theta}}_{c1}$) [4]. As mentioned above, the solution of $\boldsymbol{\theta}_c$ is dependent on the covariance matrix $\boldsymbol{\Sigma}$. To clarify this, consider Figure 4.3c for the parameters θ_1 and θ_2, which are subject to the order constraint $H_5 : \theta_1 \geq \theta_2$. The solid contour lines show the solution of $\tilde{\boldsymbol{\theta}}_c$ if $\boldsymbol{\Sigma}$ is an identity matrix (here $\tilde{\boldsymbol{\theta}}_{c1}$) and the dot-dashed contour lines show the solution (here $\tilde{\boldsymbol{\theta}}_{c2}$) of $\tilde{\boldsymbol{\theta}}_c$ if the off-diagonal elements of $\boldsymbol{\Sigma}$ equal 0.1. In Algorithm 2, we show how the above steps are implemented in R (R Development Core Team, 2016) in

[4] Calculating restricted least squares estimates $\tilde{\boldsymbol{\theta}}$ under the assumption of a closed convex cone is a well-studied problem (Nocedal & Wright, 2006). Unfortunately, maximizing the likelihood for the complement of a closed convex cone is often not a convex optimization problem and may have multiple local optima. Therefore, we cannot compute the restricted estimates directly. We need to go through all q_1 possibilities and select the boundary solution that results in the highest log-likelihood value.

pseudo code.

To compute the penalty term value for the complement, PT_c, two parts are needed, namely the probability that the order-restricted estimates are in agreement with the complement and the number of free parameters in the complement. Both will be explained next. Reconsider Figure 4.2a and also assume again that $\boldsymbol{\Sigma} = \mathbf{I}$. The complement of $H_4 : \theta_1 \geq 0,\ \theta_2 \geq 0$, that is, H_c, is constructed by the quadrants Q_2, Q_3, and Q_4, and can be written as

$$H_c : \begin{array}{ll} \theta_1 \leq 0 \ \& \ \theta_2 \geq 0 & (Q_2) \\ \text{and} & \\ \theta_1 \leq 0 \ \& \ \theta_2 \leq 0 & (Q_3) \\ \text{and} & \\ \theta_1 \geq 0 \ \& \ \theta_2 \leq 0 & (Q_4). \end{array} \tag{4.5}$$

In this case the complement can be written out easily but, for many hypotheses, it is a difficult or even impossible task to write up the complement. The constraints in Equation 4.5 show that if an estimate is not in agreement with H_4 (i.e., does not lay in Q_1), then the estimate is automatically part of the complement of H_4. Thus, the probability (under $\boldsymbol{R\theta} = \mathbf{r}$) that both estimates are in agreement with the complement (i.e., the probability that $j^c = 2$, that is, LP_2^c) equals one minus the probability that the estimates lay in Q_1, that is, 1 - 0.25 = 0.75. If the estimates lay in Q_1, then it is per definition impossible that the estimates are in agreement with H_c and $j^c = 0$ and $LP_0^c = 0.25$. In other words, the mle's are either completely in agreement with H_c or completely *not* in agreement with H_c (i.e., $LP_1^c = 0$), it is impossible that one parameter is in agreement with H_c and that the other parameter is not in agreement with H_c as is the case when evaluating H_m. To determine the number of free parameters in the complement, first, note that $H_4 : \theta_1 \geq 0,\ \theta_2 \geq 0$ (see Figure 4.2a) does not have any free parameters and equalities but that there is one free parameter (and no equality) in $H_5 : \theta_1 \geq \theta_2$ (see Figure 4.2b). In cases where there are $F \geq 1$ free parameters and/or $q_2 \geq 1$ equalities in H_m, we have to account for these in PT_c. Notable, in general in H_m, there are q_1 restrictions that can be active or inactive and q_2 restrictions that are active (since they are equality restrictions). Therefore, in case of p parameters, there are $F = p - q_1 - q_2 = p - q$ free parameters in H_m, see Table 4.1 for examples. These will remain free in H_c, as can be deduced from the example in Figure 4.2b: The complement of H_5,

that is, $H_c : \theta_1 \leq \theta_2$ is also a closed convex cone and has also one free parameters following from the reasonings used for H_m. Furthermore, the q_2 equalities in H_m will become free parameters in H_c. Hence, there are $F^c = F + q_2 = p - q_1$ free parameters in H_c. These have to be taken into account (with probability one) when computing PT_c. Additionally, there are q_1 inequality constraints in H_m that can be active or inactive. As briefly discussed before, in H_c there are either q_1 active (i.e., 0 inactive) constraints with probability LP_0^c or q_1 inactive constraints with probability $LP_{q_1}^c = 1 - LP_0^c$; and the latter equals $1 - LP_{F+q_1} = 1 - LP_{p-q_2}$ [5]. Hence, the penalty term value for the complement is defined as

$$\begin{aligned} \mathrm{PT}_c &= 1 + LP_0^c \times 0 + LP_{q_1}^c \times q_1 + 1 \times F^c \\ &= 1 + (1 - LP_{p-q_2}) \times q_1 + (p - q_1) \\ &= 1 + p - LP_{p-q_2} \times q_1. \end{aligned} \tag{4.6}$$

Note that PT_c also reduces to PT_u if there are no inequalities (i.e., $q_1 = 0$) and $q_2 \leq p$ equalities in H_m, that is, $H_m : \theta_1 = \ldots = \theta_{q_2}, \theta_{q_2+1}, \ldots, \theta_p$ and thus $H_c = H_u$. Then, $LP_{p-q_2} = 1$ and $\mathrm{PT}_c = 1 + p - LP_{p-q_2} \times 0 = 1 + p = \mathrm{PT}_u$. The penalty term value for the complement of $H_4 : \theta_1 \geq 0,\ \theta_2 \geq 0$ is computed as $\mathrm{PT}_c = 1 + 2 - 0.25 \times 2 = 2.5$ and the penalty term value for the complement of $H_5 : \theta_1 \geq \theta_2$ is computed as $\mathrm{PT}_c = 1 + 2 - 0.5 \times 1 = 2.5$. The latter is a closed convex cone and as is to be expected PT_m for the complement of H_5 and equals 2.5 like the PT_c does. In Appendix J, we illustrate the computation of the PT_m and the PT_c values in case of three parameters.

[5] If there are $F \geq 0$ free parameters in H_m, then the first F level probabilities of H_m, that is, LP_0 to LP_{F-1}, are zero. When there are q_1 inequalities in H_m, the $q_1 + 1$ level probabilities LP_F to LP_{F+q_1} sum to 1 (when $q_1 = 0$ this reduces to $LP_F = 1$, also for $F = 0$). In case of q_2 equalities (and thus q_2 constant parameters), the last q_2 level probabilities of H_m, that is, LP_{p-q_2+1} to LP_p, are zero (stated otherwise, LP_{F+q_1+1} to $LP_{F+q_1+q_2}$ are zero). Thus, the probability that there are q_1 inactive constraints in H_m (together with F free parameters and q_2 active constraints due to equalities, and in total thus $p = F + q_1 + q_2$ parameters) is $LP_{F+q_1} = LP_{p-q_2}$, where $p - q_2$ is the number of non-constant parameters in H_m. Consequently, the probability that there are q_1 inactive constraints in H_c (the complement of H_m) is $LP_{q_1}^c = 1 - LP_{F+q_1} = 1 - LP_{p-q_2}$.

4.3.1 GORIC weights

Once the GORIC values are known, the GORIC weights can be easily obtained as follows

$$w_s = \frac{\exp\{-0.5(\text{GORIC}_s)\}}{\exp\{-0.5(\text{GORIC}_m)\} + \exp\{-0.5(\text{GORIC}_c)\}}, \tag{4.7}$$

where the subscript s equals m or c for hypothesis H_m and hypothesis H_c, respectively. From these weights, we can determine the relative evidence for H_m against its complement w_m/w_c. This ratio is interpreted as the weight of evidence for H_m given the data and H_c (Kuiper, 2011, p. 106). For example, for Figure 4.3c with $\mathbf{\Sigma} \neq \mathbf{I}$, $n = 50$ and $f = 0.20$, the relative evidence for $H_5 : \theta_1 \geq \theta_2$ compared to H_c equals $w_5/w_c = 0.92/0.08 = 11.50$. This means that the order-constrained hypothesis H_5 is 11.50 times more likely than its complement. To contrast, if we want to determine the evidence ratio for H_5 against the unconstrained hypothesis H_u, that is, w_5/w_u, we have to replace the GORIC_c by the GORIC_u in Equation 4.7 and s equals m or u for hypothesis H_m and hypothesis H_u, respectively. Note that w_5 now not equates the w_5 from above, since the weights depend on the set of hypotheses. Therefore, if H_c is replaced by H_u, the weights must be recomputed for the two hypotheses in the set. Then, the evidence ratio equals $w_5/w_u = 0.62/0.38 \approx 1.63$. This clearly shows the advantage of using the complement as competing hypothesis. Next, we investigate the performance of these evidence ratio weights by means of a simulation study.

4.4 Simulation study

4.4.1 Design

We generated 500 samples according to the ANOVA model [6] $y_i = \mu_1 x_{i1} + \ldots + \mu_p x_{ip} + \epsilon_i,\ i = 1, \ldots, n$, where we assume that the residuals are normally distributed. We considered the order-constrained hypothesis $H_1 : \mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4$, its complement $H_c : \neg H_1$ and the unconstrained hypothesis $H_u : \mu_1\ , \mu_2\ , \mu_3\ , \mu_4$. Note that H_c does not equal

[6] Note that the ANOVA model is a special case of the multiple regression model discussed in the previous section.

$\mu_1 \geq \mu_2 \geq \mu_3 \geq \mu_4$; it does contain this but also the other (22) orderings of combinations of μ_1 to μ_4 (excluding the one ordering in H_m). Data were generated under hypothesis H_1 with uncorrelated independent means, for $p = 4$ groups of size $n = 30, 50, 100, 200, 500$ per group and for a variety of differences among the population means, using effect-size $f = 0, 0.10, 0.20, \ldots, 1$. Notably, $f = 0$ corresponds to sampling from the boundary of both H_m and H_c. If we sample values from a H_1 population with increasing effect-size, this will evidently lead to more and more support for H_1. Let the differences between the means, d, be equally spaced, where d is defined as $d = \frac{2f\sqrt{p}}{\sqrt{\sum_{i=1}^{p}(2i-1-p)^2}}$ under the restriction that $\sum_{i=1}^{p} \mu_i = 0$ and $\sigma = 1$. Then, the p ordered means can be computed as $\mu_i = \frac{-(p-1)d}{2} + (i-1)d$. Table 4.2 shows the computed population means for the various effect-sizes (f). The GORIC and the related weights are obtained using the procedure discussed in the previous section. The R code to run the simulations is given in Appendix K.

4.4.2 Results

All results are obtained using the `R` package restriktor (see `http://www.restriktor.org`) employing the `goric` function. The results of the simulation study are presented in Figures 4.4, 4.5, 4.6 and 4.7 and are obtained by computing the mean value of the relative evidences in each of the 500 simulation runs. Furthermore, to improve the visibility we took the natural logarithm values of the means and the range of sample-sizes and effect-sizes may vary in the figures.

The results clearly illustrate the benefits of evaluating H_m versus its complement. The mean relative evidence for H_1 versus H_c (mean w_1/w_c) increases rapidly for larger effect-sizes (see Figures 4.4a, 4.4b, 4.4c and 4.4d) and sample-sizes (see Figures 4.5a, 4.5b, 4.5c and 4.5d), while the mean relative evidence using the unconstrained hypothesis as competing hypothesis (mean w_1/w_u) is clearly bounded after a certain value. To illustrate, consider for example Figure 4.4c, where the mean relative evidence for H_1 versus H_c (mean w_1/w_c) for a medium effect-size ($f = 0.30$) is $\exp(2.63) \approx 13.87$ (on the original scale), while the mean relative evidence for H_1 versus H_u (mean w_1/w_u) is bounded at $\exp(1.92) \approx 6.82$. The value 1.92 equals the difference in penalty term values; with PT_u -

$\mathrm{PT}_m = (1.00 + 4.00) - (1.00 + 2.08) \approx 1.92$, since the log-likelihood values are here the same (i.e., $\mathrm{LL}_u = \mathrm{LL}_m$).

For small effect-sizes and small samples (see Figures 4.4a and 4.5a), the complement is slightly lower than the unconstrained hypothesis. For example, for $f = 0.10$ and $n = 30$ the mean relative evidence for w_1/w_c is $\exp(1.50) \approx 4.48$ and for w_1/w_u the mean relative evidence is $\exp(1.61) \approx 5.00$. In this case, using the complement is a bit more conservative; although the conclusion is not different of course. Furthermore, the relative evidence for small effect-sizes ($f \leq 0.20$) does not increase very rapidly (see Figures 4.4, 4.5a and 4.5b), independent of sample-size. This is because, when examining small effects using small sample-sizes, the complement is often true (even though the data were generated under H_1). This is illustrated in Figure 4.6. For example, if $f = 0$, the mle's are (except from some sampling variation) in 23/24 (approximately 95.8%) of the time not in agreement with H_1 (and thus in agreement with H_c). Thus, both hypotheses H_c and H_u have the same maximized log-likelihood value with a probability of $\mathrm{prob}_{cu} = 23/24$. When f increases, the data / the mle's will be more and more in agreement with H_1, and thus not with its complement H_c and hence the proportion of equal maximized log-likelihood values of H_u and H_c (and thus prob_{cu}) decreases. Logically, the proportion of equal maximized log-likelihood values of H_1 and H_u, that is 1 - prob_{cu}, then increases.

In Figure 4.7, the results are shown for the situation that the complement is true. Data were generated under the complement of H_1, for which we choose $H_c : \mu_1 \geq \mu_2 \geq \mu_3 \geq \mu_4$ and for the means given in Table 4.2. Note that the means are now in reversed order compared with the previous simulation. Again, we considered the order-constrained hypothesis H_1, its complement H_c and the unconstrained hypothesis H_u. The results in Figure 4.7a and 4.7b show that the mean relative evidence for H_1 versus H_c (mean w_1/w_c) and for H_1 versus H_u (mean w_1/w_u) decreases rapidly for larger f. This is because, when the effect-size and/or the sample-size increases, the data / mle's will be more and more in agreement with the complement H_c and therewith also with the unconstrained hypothesis H_u. The results shown in Figure 4.7c and 4.7d are based on the same numerical results shown in Figure 4.7a and 4.7b but now for H_c versus H_1 (mean w_c/w_1) and for H_u versus H_1 (mean w_u/w_1). They clearly show the nice property that if the complement (and also H_u) is

true, both evidence ratios w_c/w_1 and w_u/w_1 are boundless.

4.4.3 Conclusion

The results show the benefits of evaluating an order-constrained hypothesis against its complement. While, for small effect-sizes and/or sample-sizes, the difference between the evidence ratio for the true H_m when using the complement as competing hypothesis is minimal, the difference increases rapidly and profoundly for larger effect-sizes and/or sample-sizes. More importantly, the evidence ratio for the true H_m against its complement is boundless for increasing effect-sizes and/or sample-size, whereas when using the unconstrained hypothesis as competing hypothesis the evidence ratio has an upper bound. Therefore, in case that the unconstrained hypothesis is used as competing hypothesis, we recommend to replace it by the complement of the hypothesis of interest. In the next section, the method is illustrated using an empirical example about facial burn injury.

4.5 Burns example

To illustrate the method, we analyzed an empirical sample in which we sought to determine possible risk factors for ruminating thoughts after a burn injury. The data are based on a cohort study consisting of 245 individuals with burns, aged 18 to 74 years old. The response variable is rumination. Moreover, for the current illustration, we included gender (0 = men, 1 = women) and facial burns (0 = no, 1 = yes) together with its interaction as predictor variables and Hospital Anxiety and Depression Scale (HADS; Mean = 3.85, SD = 3.66), age (Mean = 41.06, SD = 13.94) and the number of surgical operations, which is a measure of severity of the burns (SO; Mean = 1.14, SD = 1.76) as covariates.

A burn event can have an avers impact on a person's quality of life. The scars can affect physical appearance and may constitute a source of rumination acting as a reminder to the event. The aim of this study example was to investigate factors that may enhance or maintain ruminating thoughts, a central concept related to depression (Nolen-Hoeksema, Wisco, & Lyubomirsky, 2008) with special interest in the role of facial burns as a risk factor for rumination. Evidence is emerging that

environmental characteristics may also contribute to the activation or maintenance of rumination. For example, in an earlier study in patients with burns a relationship between burn severity and rumination was observed (Van Loey et al., 2013). This suggests that there may be environmental triggers that influence how people cope with an adverse event. Therefore, it is expected that injury characteristics that may be perceived as distressing such as facial burn injury and larger burns may be triggers for the activation and prolongation of rumination. In addition, a gender effect is also expected because disfiguring scars resulting from burns may be of greater importance to woman as compared to men (Ghriwati et al., 2017). Then, the hypothesis of interest is $H_6 : \{\mu^{\text{adj}}_{\text{men; no facial burns}}, \mu^{\text{adj}}_{\text{men; facial burns}}, \mu^{\text{adj}}_{\text{women; no facial burns}}\} \leq \mu^{\text{adj}}_{\text{women; facial burns}}$, where μ^{adj} are the population means for rumination for the four groups determined by gender and facial burns, adjusted for the population effects of the covariates. This order-constrained hypothesis states that the means of rumination for men with and without facial burn injury and the mean of rumination for women without facial burn injury would be lower than the mean of rumination for women with facial burn injury. Note that no particular order is assumed among the first three means.

A natural choice to evaluate the order-constrained hypothesis H_6 would be an order-restricted 2×2 ANCOVA model. Since an ANCOVA is just a special case of the linear regression model, the model can be written as a linear function. To obtain adjusted means for a person with an average score on the covariates, the covariates HADS, age and SO are centered at their average and are denoted by Z_HADS, Z_age and Z_SO, respectively. Then, the model can be written as follows:

$$\begin{aligned}\text{Rumination}_i = \theta_1 &+ \theta_2 \text{facialBurns}_i + \theta_3 \text{gender}_i + \theta_4 \text{gender}_i \times \text{facialBurns}_i \\ &+ \theta_5 \text{Z\_HADS}_i + \theta_6 \text{Z\_age}_i + \theta_7 \text{Z\_SO}_i + \epsilon_i, \\ &\text{where } i = 1, \ldots, 245.\end{aligned}$$

On the left-hand side of the = operator, we have the response variable rumination and on the right-hand side we have the factors facial burns and gender and its interaction, and the centered covariates Z_HADS, Z_age and Z_SO. The interaction between gender and facial burns is included using the $\times$ operator. Then, the four adjusted means with average scores

on the covariates are computed as:

$$\begin{aligned}
\mu^{\text{adj}}_{\text{men; no facial burns}} &= \theta_1 \\
\mu^{\text{adj}}_{\text{men; facial burns}} &= \theta_1 + \theta_2 \\
\mu^{\text{adj}}_{\text{women; no facial burns}} &= \theta_1 + \theta_3 \\
\mu^{\text{adj}}_{\text{women; facial burns}} &= \theta_1 + \theta_2 + \theta_3 + \theta_4.
\end{aligned}$$

Next, we show in 7 steps how to compute the relative evidence for hypothesis H_6 compared to its complement. Again, we use the restriktor package for the analysis.

Step 1: Load your data set into `R`.

```
burns <- read.csv("burns.csv", header = TRUE, sep = " ")
```

More information about how to get your data into `R`, can be found online at `http://restriktor.org/tutorial/importdata.html`.

Step 2: Center the covariates HADS, age and SO at their average. This can be done in `R` as follows:

```
burns$Z_HADS <- burns$HADS - mean(burns$HADS, na.rm = TRUE)
burns$Z_age  <- burns$age  - mean(burns$age,  na.rm = TRUE)
burns$Z_SO   <- burns$SO   - mean(burns$SO,   na.rm = TRUE)
```

Step 3: Fit the unconstrained linear regression model using the `lm()` function.

```
fit.lm <- lm(Rumination ~ 1 + gender + facialBurns +
                              gender:facialBurns +
                              Z_HADS + Z_age + Z_SO,
                              data = burns)
```

For clarity reasons, we explicitly added an intercept term by specifying the value 1. The interaction between gender and facial burns is included using the `:` operator.

Step 4: Create the constraint syntax for **restriktor**. Now that the model is defined in R, we are left with specifying the order constraints. This

is done in restriktor by specifying a so-called constraint syntax. Order constraints are defined by means of inequality constraints (`<` or `>`) or by equality constraints (`==`). In addition, a convenient feature of the restriktor constraint syntax is the option to define new parameters that are linear in the original model parameters. This can be done using the `:=` operator. In this way, we can compute the four adjusted means and impose order constraints among these means. The constraint syntax is enclosed within single quotes. Then, for hypothesis H_6 the constraint syntax might looks as follows:

```
myConstraints <- '
  m1 := .Intercept.
  m2 := .Intercept. + facialBurns
  m3 := .Intercept. + gender
  m4 := .Intercept. + facialBurns + gender +
                      gender.facialBurns

  m1 < m4
  m2 < m4
  m3 < m4 '
```

It is important to note that variable/factor names of the interaction effects in objects of class `lm` contain a semi-colon (`:`) between the variable names (e.g., `gender:facialBurns`). To use these parameters in the constraint syntax, the semi-colon must be replaced by a dot (`.`) (e.g., `gender.facialBurns`). In addition, the intercept of a fitted objects of class `lm` is denoted in the output as `(Intercept)` and not as 1 anymore. To use the intercept in the constraint syntax, the parentheses must also be replaced by a dot (i.e., `.Intercept.`). More information about the constraint syntax can be found online at `http://restriktor.org/tutorial/syntax.html`.

Step 5: Fit the restricted linear model using the `restriktor()` function.

```
H1.restr <- restriktor(fit.lm,
                       constraints = myConstraints)
```

The first argument to the `restriktor()` function is the fitted unconstrained `lm` object from Step 3 (`fit.lm`). The second argument is the constraint syntax created in Step 4 (`myConstraints`).

Step 6: Compute the GORIC weights and the relative evidence using the `goric()` function.

```
goric(H1.restr, complement = TRUE)
```

The first argument to the `goric()` function is the fitted object of class restriktor (`H1.restr`). To compare H_6 with its complement H_c, the argument `complement` has to be set to `TRUE` (by default it is set to `FALSE`).

Step 7: Interpret the results.

```
        model      loglik  penalty       goric  goric.weights
1   Hm1.restr  -660.02415  6.98673  1334.02176        0.89130
2  complement  -661.94222  7.17279  1338.23002        0.10870
The order-restricted hypothesis Hm1.restr is 8.200 times more
likely than its complement.
```

The results show that the order-constrained hypothesis H_6 is 0.89/-0.11 $\approx$ 8.20 times more likely than its complement. For comparison, the results for the unconstrained hypothesis (not shown here) show that hypothesis H_6 is only 0.73/0.27 $\approx$ 2.70 times more likely than the unconstrained hypothesis.

In the example the sample-size equals $n = 245$ and the effect-size is approximately equal to $f = 0.10$ but to investigate the overall performance we ran an extra simulation study for $H_{6a} : \{\mu_1, \mu_2, \mu_3\} \leq \mu_4$ with $n = 30$ and $n = 250$, and effect-sizes ranging from $f = 0$ to $f = 0.6$. The results are presented in Appendix L. The results are comparable with the simulation results shown in Figure 4.4. The mean relative evidence for H_{6a} against H_c is boundless, and the mean relative evidence for H_{6a} against H_u stabilizes at an upper-bound from a certain effect-size and sample-size (the latter is not shown here). The main difference is that, the mean relative evidence for H_{6a} against H_c increases more rapidly for smaller effect-sizes, then for H_1 against its complement. At last, if we compare the result from the example with the simulation results ($n = 250$

and $f = 0.10$) shown in Figure L.1b, then we can see that the relative evidence for $w_6/w_c \approx 8.20$ is close to the simulation results $w_{6a}/w_c \approx 9.97$ (on the original scale).

4.6 Summary and recommendations

In this paper, we introduced the evaluation of an order-constrained hypothesis against its complement using the GORIC (weights). The GORIC is an information criterion that can be used to evaluate competing hypotheses in univariate and multivariate normal linear models, where the regression parameters are subject to inequality constraints of the type $\boldsymbol{R\theta} \geq \boldsymbol{r}$, where $\boldsymbol{R}$ is a matrix with known constants, $\boldsymbol{\theta}$ a vector with the regression parameters and $\boldsymbol{r}$ a vector with known constants. The interpretation can be improved by computing GORIC weights and related evidence ratios reflecting relative evidence for one hypothesis versus another.

We advise that one should evaluate their theory against its complement H_c instead of the unconstrained hypothesis H_u. The advantage of our method is that the relative evidence for an order-constrained hypothesis H_m compared to its complement is boundless, whereas the relative evidence for H_m compared to H_u is nor increased by a larger sample-size neither by a larger effect-size, if the data are in agreement with the hypothesis of interest (i.e., theory). In a small simulation study, we showed for a true $H_1 : \mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4$ versus H_c that the mean relative evidence increases for larger sample-sizes and effect-sizes, while the relative evidence for a true H_1 versus H_u remains bounded. The method was illustrated using an empirical example about facial burn injury. In seven easy steps, we showed how to compute the relative evidence of the researchers theory against its complement using the `R` package **restriktor**. The results show that using the complement as competing hypothesis lead to much more support for the hypothesis of interest when it is true, compared to using the unconstrained hypothesis as competing hypothesis.

We assumed that researchers often do not have specific competing hypotheses. While, this is probably often the case, it is conceivable that the set of hypotheses contains more than one competing hypothesis. In these cases, the problem that the relative evidence for H_m against H_u is not

affected by increasing sample-size and/or effect-size after a specific value can still occur. For example, consider the set with three hypotheses, H_1, $H_2 : \mu_1 \leq \mu_2 \leq \mu_3 = \mu_4$ (which is a subset of H_1) and the unconstrained hypothesis H_u. If H_2 is true, then all three hypotheses are true and all evidence ratios are bounded (Kuiper et al., 2011, p. 107). However, further research is needed to investigate the evaluation of a set of multiple order-constrained hypotheses against its complement because determining the complement might not always be trivial (especially for software).

The results presented in this article are for the univariate linear regression model but fortunately they can easily be adapted for the multivariate normal linear model. One should keep in mind that, unlike in the univariate setting, where $\tilde{\boldsymbol{\theta}}$ does *not* depend on the order-restricted covariance matrix, denoted by $\tilde{\boldsymbol{\Sigma}}$, in the multivariate normal linear model $\tilde{\boldsymbol{\theta}}$ does depend on $\tilde{\boldsymbol{\Sigma}}$ and $\tilde{\boldsymbol{\Sigma}}$ on $\tilde{\boldsymbol{\theta}}$ Kuiper, Hoijtink, and Silvapulle (2012). Hence, an iterative procedure is needed to calculate them. The procedure is implemented in **restriktor**.

Acknowledgments

The first author is a PhD fellow of the research foundation Flanders (FWO) at Ghent university (Belgium) and at Utrecht University (The Netherlands). The third author is supported by a grant from the Netherlands organization for scientific research: NWO-VENI-451-16-019.

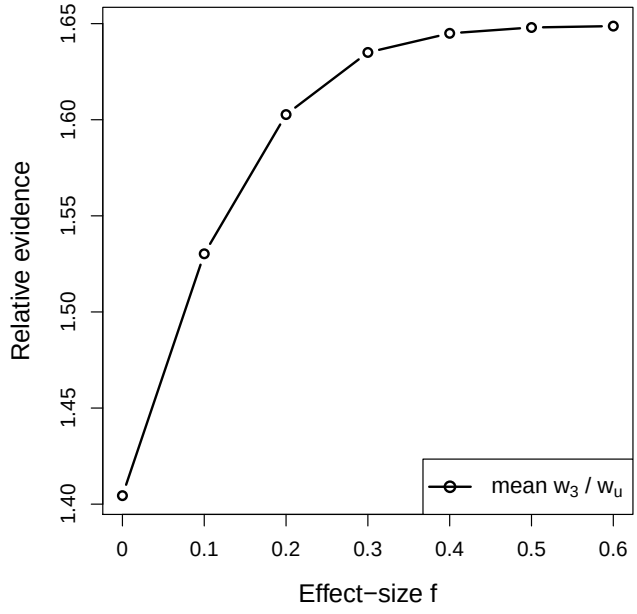
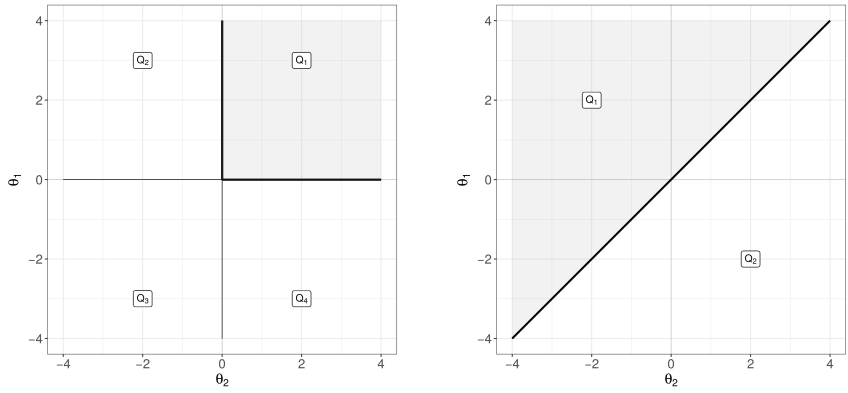
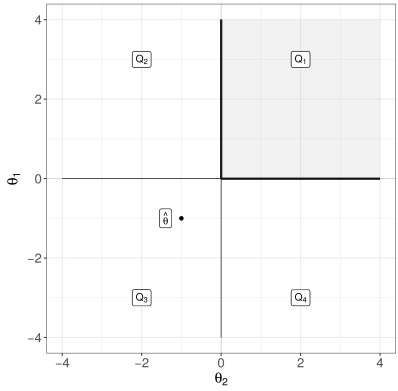


Figure 4.1: Mean relative evidence for hypothesis $H_3 : \mu_1 \geq \mu_2$ compared to the unconstrained hypothesis (mean w_3/w_u) when H_3 is true, for $n = 50$ and various effect-sizes (f).

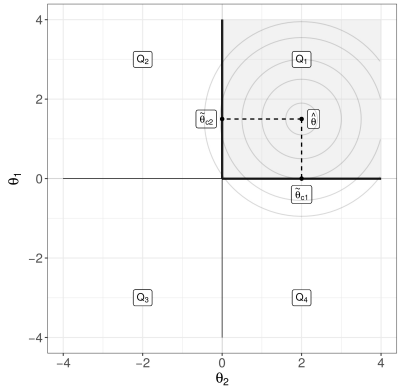


(a) $H_4 : \theta_1 \geq 0,\ \theta_2 \geq 0.$ (b) $H_5 : \theta_1 \geq \theta_2.$

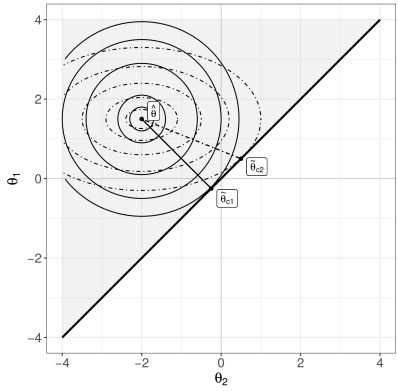
Figure 4.2: Illustration to illuminate on the computation of the penalty term value of H_m. The gray-shaded area is the permissible area under H_m.



(a) $H_4 : \theta_1 \geq 0,\ \theta_2 \geq 0$. The mle's $\hat{\boldsymbol{\theta}}$ lay in Q_3 and is thus in agreement with H_c.



(b) $H_4 : \theta_1 \geq 0,\ \theta_2 \geq 0$, for $\boldsymbol{\Sigma} = \mathbf{I}$. The mle's $\hat{\boldsymbol{\theta}}$ lay in Q_1 and is thus *not* in agreement with H_c.



(c) $H_5 : \theta_1 \geq \theta_2$, for $\boldsymbol{\Sigma} = \mathbf{I}$ (solid line) and $\boldsymbol{\Sigma} \neq \mathbf{I}$ (dot-dashed line).

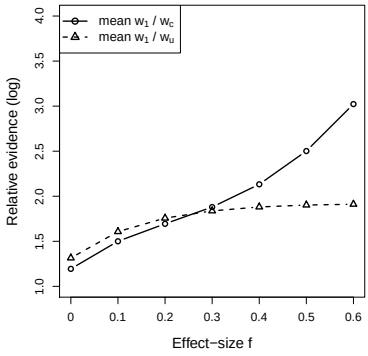
Figure 4.3: The gray-shaded area is the permissible area under H_m.

Algorithm 2 Compute the log-likelihood$_c$ value.

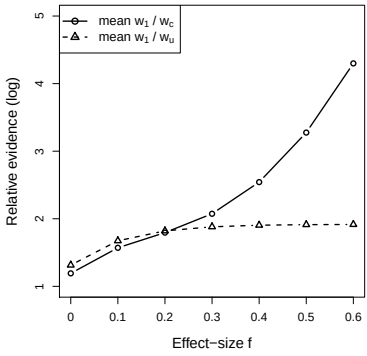
if not all ($\boldsymbol{R}_1\hat{\boldsymbol{\theta}}$ - $\mathbf{r}_1 \geq \mathbf{0}$) and/or not all ($\boldsymbol{R}_2\hat{\boldsymbol{\theta}}$ - $\mathbf{r}_2 = \mathbf{0}$) **then** ▷ Check if any constraint is violated.
  return log-likelihood$_u$ ▷ LL$_c$ = LL$_u$
else ▷ Note that equality constraints are freed. Hence, we only use the constraint matrix $\boldsymbol{R}_1$.
  nr ⇐ 1 to nrow($\boldsymbol{R}_1$) ▷ Vector from 1 to the number of rows of $\boldsymbol{R}_1$.
  q_1 ⇐ length(nr) ▷ Length nr vector: q_1.
  for b ← 1 to q_1 **do**
    idx ⇐ vector(nr[b], nr[-b])
    $\boldsymbol{R}_1$.idx ⇐ $\boldsymbol{R}_1$[idx,] ▷ Put row b of matrix $\boldsymbol{R}_1$ on top.
    LL ⇐ RESTRIKTOR(model, constraints = $\boldsymbol{R}_1$.idx, neq = 1)
      ▷ The first row of $\boldsymbol{R}_1$.idx is treated as equality constraints.
    log-likelihood[b] ⇐ LL ▷ Store the LL value at the bth position of the log-likelihood vector.
  end for
  log-likelihood$_c$ ⇐ max(log-likelihood) ▷ Select the highest value in the log-likelihood vector for the log-likelihood$_c$ value.
  return log-likelihood$_c$
end if

Table 4.1: Examples of the number of free parameters F for various order-constrained hypotheses H_m.

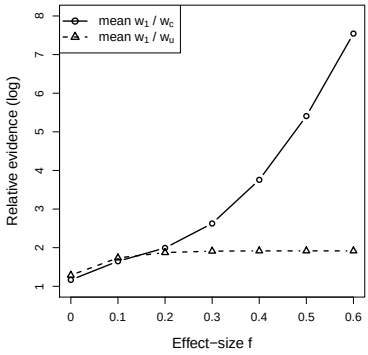
H_m	Number of free parameters $F_m = p - q_1 - q_2$	Comments
$H_a : \theta_1 \leq \theta_2 \leq \theta_3$	$F_a = 3 - 2 - 0 = 1$	Fix one parameter and the other two are bounded.
$H_b : \theta_1 \leq \theta_2 + \theta_3$	$F_b = 3 - 1 - 0 = 2$	Fix one and the others are still free. Fix two and the other one is bounded.
$H_c : \theta_1 \leq \theta_2 + \theta_3, \theta_4$	$F_c = 4 - 1 - 0 = 3$	Additional free parameter θ_4, thus $F_c = F_b + 1$.
$H_d : \theta_1 \leq \theta_2 + \theta_3, \theta_4 = 0.5$	$F_d = 4 - 1 - 1 = 2$	Additional constant parameter θ_4, thus $F_d = F_b + 0$.
$H_e : \theta_1 \leq 0.5, \theta_2 \leq 0.5$	$F_e = 2 - 2 - 0 = 0$	Both parameters are bounded.
$H_f : \theta_1 \leq \theta_2, \theta_2 \leq 0.5$	$F_f = 2 - 2 - 0 = 0$	Both parameters are bounded.
$H_g : \theta_1 \leq 0.5, \theta_2 = 0$	$F_g = 2 - 1 - 1 = 0$	The first parameter is bounded and the second parameter is a constant.



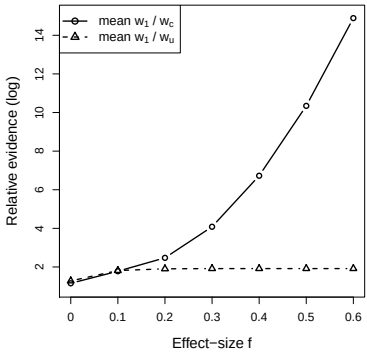
(a) $n = 30$.



(b) $n = 50$.

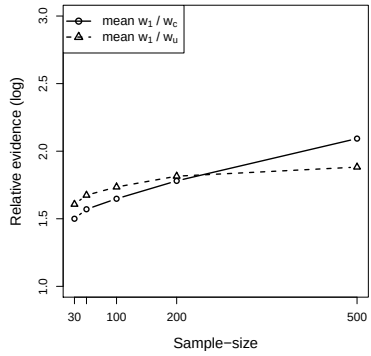


(c) $n = 100$.

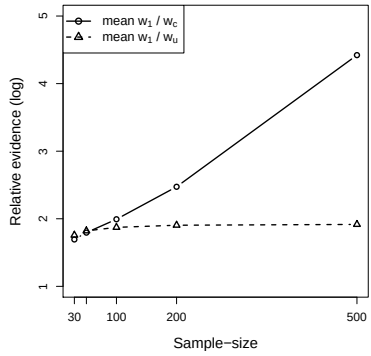


(d) $n = 200$.

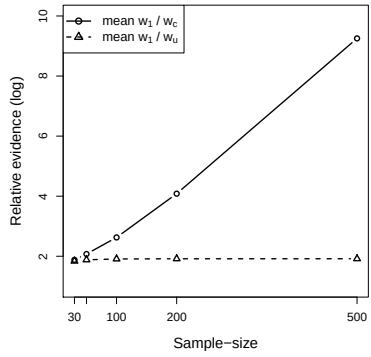
Figure 4.4: Mean of the relative evidence (on a log scale) for the situation that the order-constrained hypothesis H_1 is true. Hypothesis H_1 is compared to its complement H_c (mean w_1/w_c) and to the unconstrained hypothesis H_u (mean w_1/w_u) for various effect-sizes (f) and for $n = 30, 50, 100$ and 200.



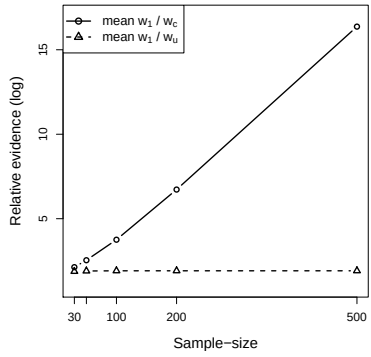
(a) $f = 0.10$.



(b) $f = 0.20$.

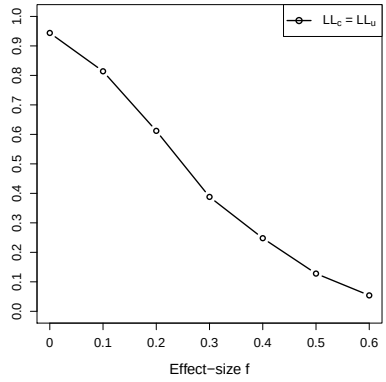


(c) $f = 0.30$.

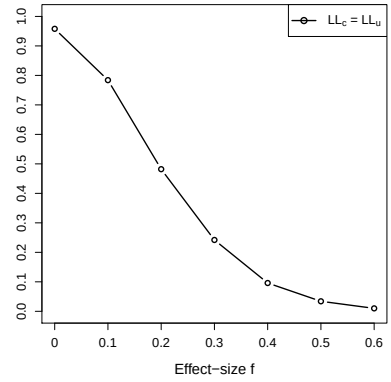


(d) $f = 0.40$.

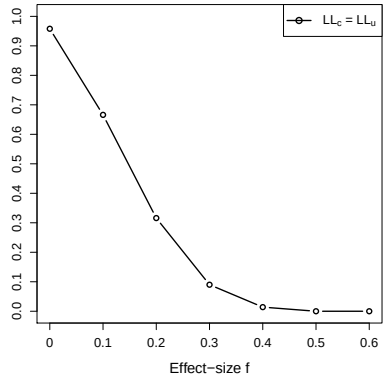
Figure 4.5: Mean of the relative evidence (on a log scale) for the situation that the order-constrained hypothesis H_1 is true. Hypothesis H_1 is compared to its complement H_c (mean w_1/w_c) and to the unconstrained hypothesis H_u (mean w_1/w_u) for various sample-sizes (n) and for $f = 0.10, 0.20, 0.30$ and 0.40.



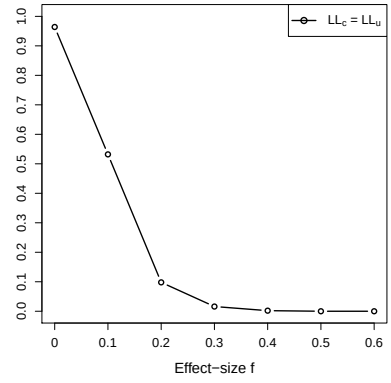
(a) $n = 30$.



(b) $n = 50$.

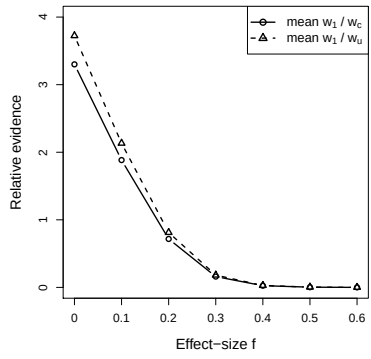


(c) $n = 100$.

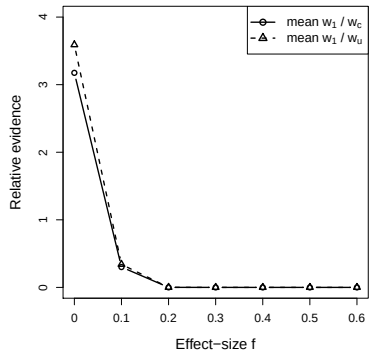


(d) $n = 200$.

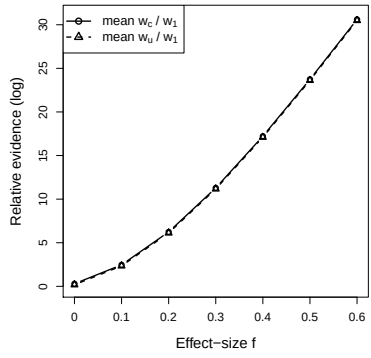
Figure 4.6: Proportion of data sets that result in equal log-likelihood values for the complement of $H_1 : \mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4$, that is, H_c, and the unconstrained hypothesis H_u (i.e., $\mathrm{LL}_c = \mathrm{LL}_u$), for various effect-sizes (f) and $n = 30, 50, 100, 200$.



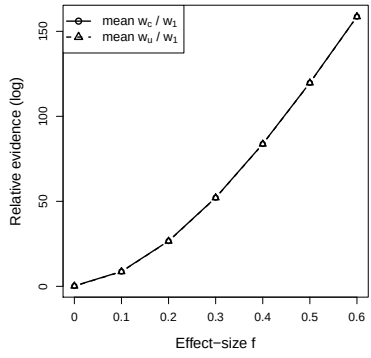
(a) $n = 30$.



(b) $n = 200$.



(c) $n = 30$.



(d) $n = 200$.

Figure 4.7: Mean of the relative evidence for the situation that the complement H_c of the order-constrained hypothesis H_1 is true, for various effect-sizes (f) and for $n = 30$ and 200. (a,b) H_1 versus H_c (mean w_1/w_c), and H_1 versus the unconstrained hypothesis H_u (mean w_1/w_u). (c,d) H_c versus H_1 (mean w_c/w_1), and H_u versus H_1 (mean w_u/w_1). Both c and d are on a log-scale.

Table 4.2: Population means for first simulation study.

Effect-size	population means			
f	μ_1	μ_2	μ_3	μ_4
0	0	0	0	0
0.1	-0.134	-0.044	0.044	0.134
0.2	-0.268	-0.089	0.089	0.268
0.3	-0.402	-0.134	0.134	0.402
⋮	⋮	⋮	⋮	⋮
1	-1.341	-0.447	0.447	1.341

Note: in the second simulation, we used the reverse ordering of these means.

J Example of computing the PT_c in case of 3 parameters

In the example of Figure $A1$, the unrestricted parameter space is determined by the three parameters θ_1, θ_2 and θ_3 (and is of course the whole space and not just the one depicted in Figure $A1$). The gray-shaded area is a closed convex cone and is defined by the order constraints $H_{A1} : \theta_1 \leq \theta_2, \theta_1 \leq \theta_3$. The penalty term value for the unconstrained hypothesis equals $\mathrm{PT}_u = 1 + p = 1 + 3 = 4$. The level probabilities corresponding to H_{A1} equal $LP_0 = 0, LP_1 = \frac{1}{6}, LP_2 = \frac{1}{2}$ and $LP_3 = \frac{1}{3}$. A level probability of $LP_0 = 0$ means that it is impossible that the vector with order-constrained estimates $\tilde{\boldsymbol{\theta}}$ has zero inactive constraints (i.e., $j = 0$). This is because, we have one free parameter $F = p - q_1 - q_2 = 3 - 2 - 0 = 1$, which is per definition inactive. The level probability $LP_3 = \frac{1}{3}$ is the probability that the vector with order-constrained estimates $\tilde{\boldsymbol{\theta}}$ is identical to the unconstrained estimates (i.e., $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}$ and $j = 3$). This probability corresponds to the proportion of the gray-shaded area compared to the whole cube in Figure $A1$ and of course also of H_m versus the whole space. Hence, the penalty term value for H_{A1}

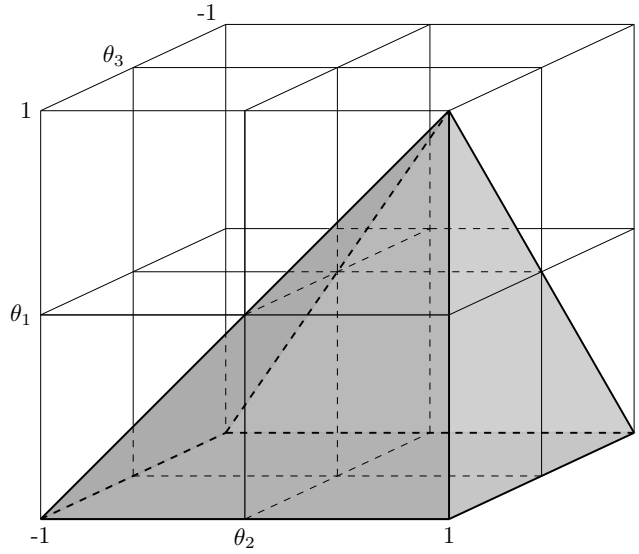


Figure J.1: The permissible gray area is defined by $H_{A1} : \theta_1 \leq \theta_2, \theta_1 \leq \theta_3$, depicted for θ_1, θ_2 and θ_3 between -1 and 1.

equals $\mathrm{PT}_{A1} = 1 + 0 \times 0 + \frac{1}{6} \times 1 + \frac{1}{2} \times 2 + \frac{1}{3} \times 3 = 3\frac{1}{6}$. For the complement H_c, which corresponds to the not gray-shaded area in the cube, the probability that there are $q_1 = 2$ inactive constraints is $LP^c_{q_1} = 1 - LP_{p+q_2}$. Since, there are $q_2 = 0$ equality constraints, the penalty term value for H_c equals $\mathrm{PT}_c = 1 + p - LP_{p-q_2} \times q_1 = 1 + 3 - \frac{1}{3} \times 2 = 3\frac{1}{3}$.

K

R-code simulation (for $n = 200$)

```
# install the restriktor package
install.packages("restriktor")

# load restriktor library
library(restriktor)

# sample-size
n <- 200
# number of parameters
p <- 4
# define constraints
R1 <- 'x1 < x2; x2 < x3; x3 < x4'
# effect-sizes
es <- seq(0,1,.1)
# identity covariance matrix
Sigma <- diag(p)
# number of simulation runs
nsim <- 500

# create list for storing the relative weights from the simulation
out.goric.wt <- list()

for (k in 1:length(es)) {
  # compute the p=4 population means
  j <- 1:p
```

```
# compute equal difference scores, d, between the means
d <- (2*sqrt(p)*es[k]) / sqrt(sum((2*j-1-p)^2))
# compute p means
means <- ((-(p-1)*d ) / 2) + (j - 1)*d

# Run the nsim = 500 simulations/iterations
for (i in 1:nsim) {
  cat("iteration =", k, "... =", i, "\n")
  set.seed(3013073 + i)

  # generate data
  y <- cbind(c(matrix(MASS:::mvrnorm(n, mu = means, Sigma = Sigma,
                                     empirical = FALSE), nrow = n)))
  # create p groups
  x <- factor(rep(1:p, each = n))

  # fit linear model
  fit.lm <- lm(y ~ -1 + x)

  # fit order-constrained model
  # to speed-up the simulations we switched off the computation of
  # the standard errors (se = "none").
  Hm.restr <- restriktor(fit.lm, constraints = R1, se = "none")
  # fit unconstrained model
  Hu.restr <- restriktor(fit.lm, se = "none")

  ## compute goric
  # Hm versus Hc
  GORICc <- restriktor:::goric(Hm.restr, complement = TRUE)
  # Hm versus Hu
  GORICu <- restriktor:::goric(Hm.restr, Hu.restr)

  # goric value Hm.
  # Note: this value is of course the same value as in GORICu$goric[1]
  goric[i,1] <- GORICc$goric[1]
  # goric value Hc
  goric[i,2] <- GORICc$goric[2]
  # goric value Hu
  goric[i,3] <- GORICu$goric[2]
}

goric.wt <- matrix(NA, nsim, 2)

# compute the goric weights for Hm versus Hc and for Hm versus Hu for each
# of the 500 data sets.
for (j in 1:nsim) {
  goric.Hm <- goric[j,1]
  goric.Hc <- goric[j,2]
  goric.Hu <- goric[j,3]
  delta.Hc <- c(goric.Hm, goric.Hc) - min(c(goric.Hm, goric.Hc))
```

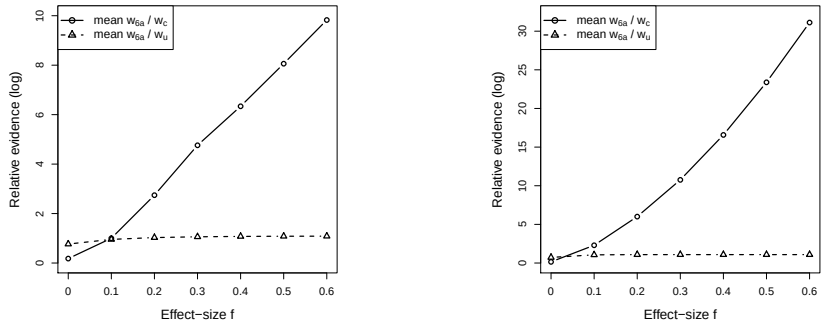
```
    delta.Hu <- c(goric.Hm, goric.Hu) - min(c(goric.Hm, goric.Hu))
    goric.weights.Hc <- exp(-delta.Hc / 2) / sum(exp(-delta.Hc / 2))
    goric.weights.Hu <- exp(-delta.Hu / 2) / sum(exp(-delta.Hu / 2))
    goric.wt[j,1] <- goric.weights.Hc[1] / goric.weights.Hc[2]
    goric.wt[j,2] <- goric.weights.Hu[1] / goric.weights.Hu[2]
  }

  out.goric.wt[[k]] <- goric.wt

}
```

L

Simulation results for hypothesis H_{6a}



(a) $n = 30$. (b) $n = 250$.

Figure L.1: Mean of the relative evidence (on a log scale) for the situation that the order-constrained hypothesis H_{6a} is true. Hypothesis H_{6a} is compared to its complement H_c (mean w_{6a}/w_c) and to the unconstrained hypothesis H_u (mean w_{6a}/w_u) for various effect-sizes (f) and for $n = 30$ and 250.

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), (pp. 199–213). Springer New York: NY. doi: doi:10.1007/978-1-4612-1694-0_15

Anraku, K. (1999). An information criterion for parameters under a simple order restriction. *Biometrika*, *86*(1), 141–152. doi: doi:10.1093/biomet/86.1.141

Burnham, K., & Anderson, D. (2002). *Model selection and multimodel inference: a practical information-theoretic approach* (2nd ed.). Spring-Verlag: New York.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Ghriwati, A., Sutter, M., Pierce, B., Perrin, P., Wiechman, S., & Schneider, J. (2017). Two-year gender differences in satisfaction with appearance after burn injury and prediction of five-year depression: A latent growth curve approach. *Archives of Physical Medicine and Rehabilitation*. doi: doi:10.1016/j.apmr.2017.04.011

Grömping, U. (2010). Inference with linear equality and inequality constraints using R: The package ic.infer. *Journal of statistical software*, *33*, 1–31. doi: doi:10.18637/jss.v033.i10

Hoijtink, H. (2012). *Informative Hypotheses: Theory and Practice for Behavioral and Social Scientists*. Boca Raton, FL: Taylor & Francis.

Kuiper, R. (2011). *Model selection criteria: How to evaluate order restrictions* (Dissertation, Utrecht University). Retrieved from `https://dspace.library.uu.nl/handle/1874/224499`

Kuiper, R., Hoijtink, H., & Silvapulle, M. (2011). An akaike-type information criterion for model selection under inequality constraints. *Biometrika*, *98*(2), 495–501. doi: doi:10.1093/biomet/asr002

Kuiper, R., Hoijtink, H., & Silvapulle, M. (2012). Generalization of the order-restricted information criterion for multivariate normal linear models. *Journal of Statistical Planning and Inference*, *142*(8), 42454–2463. doi: doi:10.1016/j.jspi.2012.03.007

Nocedal, J., & Wright, S. (2006). *Numerical Optimization* (2nd ed.; V. Mikosh, S. Resnick, & S. Robinson, Eds.). Spring-Verlag: New York.

Nolen-Hoeksema, S., Wisco, B. E., & Lyubomirsky, S. (2008). Rethinking Rumination. *Perspectives on Psychological Science*, *3*, 400–24. doi: doi:10.1111/j.1745-6924.2008.00088.x

R Development Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. (ISBN 3-900051-07-0)

Silvapulle, M., & Sen, P. (2005). *Constrained statistical inference: Order, inequality, and shape restrictions.* Hoboken, NJ: Wiley.

Van Loey, N., Oggel, A., Goemanne, A., Braem, L., Vanbrabant, L., & Geenen, R. (2013). Cognitive emotion regulation strategies and neuroticism in relation to depressive symptoms following burn injury: a longitudinal study with a 2-year follow-up. *Journal of Behavioral Medicine*.

Wolak, F. (1987). An exact test for multiple inequality and equality constraints in the linear regression model. *Journal of the American statistical association*, *82*(399), 782–793. doi: doi:10.1080/01621459.1987.10478499

Software

5

A General Procedure for Testing Inequality Constrained Hypotheses in SEM[1]

Researchers in the social and behavioral sciences often have clear expectations about the order and/or the sign of the parameters in their statistical model. For example, a researcher might expect that regression coefficient β_1 is larger than β_2 and β_3. To test such a constrained hypothesis special methods have been developed. However, the existing methods for structural equation models (SEM) are complex, computationally demanding and a software routine is lacking. Therefore, in this paper we describe a general procedure for testing order/inequality constrained hypotheses in SEM using the `R` package `lavaan`. We use the likelihood ratio statistic to test constrained hypotheses and the resulting plug-in p value is computed by either parametric or Bollen-Stine bootstrapping. Since the obtained

[1]This chapter is published as Vanbrabant, L., Van de Schoot, R., Van Loey, N., & Rosseel, Y. (2017). A General Procedure for Testing Inequality Constrained Hypotheses in SEM. Methodology, 13: 61–70. DOI: 10.1027/1614-2241/a000123.

plug-in p value can be biased, a double bootstrap approach is available. The procedure is illustrated by a real-life example about the psychosocial functioning in patients with facial burn wounds.

5.1 Introduction

Structural equation modeling (SEM) software such as `lavaan` (Rosseel, 2012a) and `Mplus` (Muthen & Muthen, 2010) can be used to impose order/inequality constraints on the parameters of a statistical model. For example, there might be a hypothesis stating that regression coefficient β_1 is larger than regression coefficient β_2 and β_3, which is denoted by

$$H: \quad \beta_1 \geq \{\beta_2, \beta_3\}, \tag{5.1}$$

and is called an (order) constrained hypothesis (Barlow, Bartholomew, Bremner, & Brunk, 1972; Hoijtink, 2012; Klugkist, Laudy, & Hoijtink, 2005; Kuiper, Klugkist, & Hoijtink, 2010; Mulder, Hoijtink, & de Leeuw, 2012; Robertson, Wright, & Dykstra, 1988; Silvapulle & Sen, 2005; Van de Schoot, Hoijtink, Mulder, et al., 2011). In the literature, two methods are known for evaluating constrained hypotheses in SEM, namely the frequentist method proposed by (Van de Schoot, Hoijtink, & Deković, 2010) and the Bayesian method proposed by Van de Schoot, Hoijtink, Hallquist, and Boelen (2012). In this article, we focus on the frequentist procedure. However, the procedure is rather complex, since an abundant number of steps have to be carried out in `Mplus` and `R` (R Development Core Team, 2016). Besides, the procedure is computationally demanding, limited to the parametric bootstrap, and no software routine is available.

Therefore, in the current paper we describe the `R` function `InformativeTesting()`. We will show that the `InformativeTesting()` function is easy to use and more flexible than the procedure described in Van de Schoot et al. (2010). Moreover, the `InformativeTesting()` function has some additional features, namely the Bollen-Stine bootstrap (Bollen & Stine, 1993) for non-normal data, parallel processing to reduce computational time, an option to produce high-quality plots based on the results, and the procedure does not depend on third-party commercial software but uses the open-source package `lavaan`.

The remainder of the paper is organized as follows. First, we describe the general structural equation model and its parameters on which constraints can be imposed. Furthermore, two hypothesis tests are introduced for testing constrained hypotheses and an illustration is presented to show how theoretical expectations can be converted into a constrained hypothesis. Second, we present a procedure for testing constrained hypotheses. We introduce the parametric and Bollen-Stine bootstrap approaches and we discuss the genuine double bootstrap method. Third, an overview of the `InformativeTesting()` function is presented. We show by means of a five step procedure how to convert the statistical model and the constraints into `lavaan` syntax and we show how to set up the necessary function arguments. In addition, we describe the output of the `print()` and `plot()` methods using the results of the illustration. Finally, we make some concluding remarks.

5.2 Structural equation model with constraints

A SEM with latent variables consists of two parts, namely a structural model and a measurement model. The structural model represents the structural equations that summarize the relationships between latent variables and can be written as:

$$\boldsymbol{\eta}^{\boldsymbol{g}} = \boldsymbol{\alpha}^{\boldsymbol{g}} + \boldsymbol{B}^{\boldsymbol{g}}\boldsymbol{\eta}^{\boldsymbol{g}} + \boldsymbol{\Gamma}^{\boldsymbol{g}}\boldsymbol{x}^{\boldsymbol{g}} + \boldsymbol{\zeta}^{\boldsymbol{g}} \ , \tag{5.2}$$

where the superscript g denotes group membership and runs from $g = 1, \ldots, G$. The measurement model represents the link between the latent and observed variables and is written as

$$\boldsymbol{y}^{\boldsymbol{g}} = \boldsymbol{\nu}^{\boldsymbol{g}} + \boldsymbol{\Lambda}^{\boldsymbol{g}}\boldsymbol{\eta}^{\boldsymbol{g}} + \boldsymbol{K}^{\boldsymbol{g}}\boldsymbol{x}^{\boldsymbol{g}} + \boldsymbol{\epsilon}^{\boldsymbol{g}}, \tag{5.3}$$

where

$\boldsymbol{y}$	$\rightarrow$	$p \times 1$ vector of dependent variables.
$\boldsymbol{\eta}$	$\rightarrow$	$m \times 1$ vector of factors / latent variables.
$\boldsymbol{\nu}$ and $\boldsymbol{\alpha}$	$\rightarrow$	are vectors of intercepts.
$\boldsymbol{\Lambda}$	$\rightarrow$	$p \times m$ matrix of factor loadings.
$\boldsymbol{K}$ and $\boldsymbol{\Gamma}$	$\rightarrow$	matrices that contain slopes for exogenous covariates in the $(q \times 1)$ vector $\boldsymbol{x}$.
$\boldsymbol{B}$	$\rightarrow$	$m \times m$ vector of structural regression slopes.
$\boldsymbol{\epsilon}$ and $\boldsymbol{\zeta}$	$\rightarrow$	vector of error terms.
$\boldsymbol{\Phi}$	$\rightarrow$	$p \times p$ covariance matrix of $\boldsymbol{\epsilon}$.
$\boldsymbol{\Psi}$	$\rightarrow$	$m \times m$ covariance matrix of $\boldsymbol{\zeta}$.

Furthermore, $\boldsymbol{\epsilon}$ and $\boldsymbol{\zeta}$ are multivariate normal distributed with means zero and covariance matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ respectively.

The non-redundant free parameters of the model are collected in the parameter vector $\boldsymbol{\theta}$. Order/inequality constraints can be imposed on all [2] parameters of a structural equation model but in practice, only a subset of the free parameters are constrained. Then, let $\boldsymbol{\theta} = \{\boldsymbol{\theta^a}, \boldsymbol{\theta^b}\}$, where $\boldsymbol{\theta^a}$ includes all parameters on which we impose constraints and where $\boldsymbol{\theta^b}$ includes the remaining unconstrained parameters.

To test constrained hypotheses we consider two types of hypothesis tests, namely Type A and Type B (Silvapulle & Sen, 2005, pp. 61–62):

Type A:

$$\begin{aligned} H_{A0}: &\quad \boldsymbol{L\theta^a} = \boldsymbol{c} \\ H_{A1}: &\quad \boldsymbol{L\theta^a} \geq \boldsymbol{c}\ , \end{aligned} \tag{5.4}$$

Type B:

$$\begin{aligned} H_{B0}: &\quad \boldsymbol{L\theta^a} \geq \boldsymbol{c} \\ H_{B1}: &\quad \boldsymbol{\theta^a} \in \mathbb{R}^k\ . \end{aligned} \tag{5.5}$$

If l is the number of inequality constraints imposed on $\boldsymbol{\theta^a}$, and k the number of parameters involved, then let $\mathbf{L}$ be an $l \times k$ matrix with known constants, and $\boldsymbol{c}$ an $l \times 1$ vector with known constants (often this vector contains zeros). In hypothesis test Type A the null-hypothesis H_{A0}, in which all parameters are constrained to be equal, is tested against the

[2] In this article, we focus on imposing constraints on $\boldsymbol{B}, \boldsymbol{\Gamma}, \boldsymbol{K}, \boldsymbol{\Lambda}, \boldsymbol{\nu}, \boldsymbol{\alpha}$ and $\boldsymbol{\Phi}$.

constrained hypothesis H_{A1}. In hypothesis test Type B the constrained hypothesis H_{B0} is tested against the unconstrained model H_{B1}, which has no restrictions on $\boldsymbol{\theta}^a$. In order to find affirmative evidence for the constrained hypothesis, hypothesis test Type B plays a crucial role. Severe constraint violations result in rejecting the constrained hypothesis, since it is tested against the best fitting (i.e. unconstrained) hypothesis. Hypothesis test Type A is required to avoid false conclusions in case the inequality constraints are in fact equality constraints. In other words, hypothesis H_{A0} in test Type A should be rejected and the constrained hypothesis H_{B0} in test Type B not. If this is the case, loosely speaking this means that the constraints are in accordance with the data.

5.2.1 Illustration

To illustrate constrained hypothesis testing we use an example based on a cohort study in patients with facial burns (Hoogewerf, van Baar, Middelkoop, & van Loey, 2014). The example concerns a multiple group model with two groups, men and women. The sample consists of 77 respondents ($M_{age} = 39.95,\ SD = 14.05$) with facial burns, 78% of the respondents were men ($M_{age} = 38.96,\ SD = 13.76$) and 22% were women ($M_{age} = 44.04,\ SD = 14.81$). The aim of the study was to examine psychosocial functioning in patients with facial burn wounds. More in particular, in this part of the study the researchers wanted to test the hypothesis that the impact of burn severity on self-esteem would be higher in women compared to men after controlling for symptoms of anxiety and depression. Burn severity was measured by the total body surface area burned (TBSA) which is the percentage of partial and full thickness burns on the total body. Anxiety and depression symptoms, and self-esteem were measured using the HADS (Spinhoven et al., 1997) and the Rosenberg's self-esteem scale (Rosenberg, 1965) respectively.

Previous studies have emphasized the greater importance of appearance on self-esteem and body image in women compared to men. One study (Strahan, Wilson, Cressman, & Buote, 2006) reported that women made more upward social comparisons than men on the body domain. In the aftermath of a burn injury that can cause lifelong disfigurement, it was empirically confirmed that female patients with burns are more dissatisfied with their appearance, leading to worse psychosocial functioning

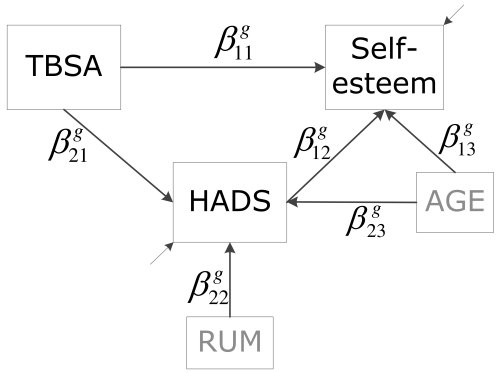


Figure 5.1: Multiple group SEM for the relation between total burned surface area (TBSA), self-esteem and symptoms of depression and anxiety (HADS) for men and women, controlling for age and coping style rumination.

(Thombs et al., 2008). Women with facial burns in particular showed to be at higher risk for long term depression symptoms (Wiechman et al., 2001). However, irrespectively of gender, depression symptoms are associated with low self-esteem and feelings of worthlessness (APA, 2000) and with maladaptive coping styles. Rumination (RUM) in particular has been strongly related to depression and anxiety symptoms (Nolen-Hoeksema, Wisco, & Lyubomirsky, 2008). The theoretical assumptions between these variables are shown in Figure 5.1.

Since we are not interested in the intercepts, we can ignore the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\nu}$ in equations 5.2 and 5.3. In the theoretical model we are dealing only with observed rather than latent variables. In the LISREL tradition, all observed variables involved in a structural equation, are upgraded to latent variables. Hence, the matrices $\boldsymbol{\Gamma}$ and $\boldsymbol{K}$ are not involved in estimating the model. Thus, we can write the model in Figure 5.1 as:

$$\begin{aligned} \boldsymbol{\eta^g} &= \boldsymbol{B^g}\boldsymbol{\eta^g} + \boldsymbol{\zeta^g} \\ \boldsymbol{y^g} &= \boldsymbol{\Lambda^g}\boldsymbol{\eta^g} + \boldsymbol{\epsilon^g} \ , \end{aligned} \tag{5.6}$$

where

$$\boldsymbol{B^g} = \begin{bmatrix} \beta_{11} & \beta_{12} & \beta_{13} \\ \beta_{21} & \beta_{22} & \beta_{23} \end{bmatrix} \ , \tag{5.7}$$

$$\boldsymbol{\Psi^g} = \begin{bmatrix} \psi_{11} & & & & \\ 0 & \psi_{22} & & & \\ 0 & 0 & \psi_{33} & & \\ 0 & 0 & \psi_{43} & \psi_{44} & \\ 0 & 0 & \psi_{53} & \psi_{54} & \psi_{55} \end{bmatrix}, \tag{5.8}$$

and $\boldsymbol{\Lambda}^g$ is an identity matrix $\boldsymbol{I}$ and $\boldsymbol{\Phi}^g = \mathbf{0}$. In our example $G = 2$, where $g = 1$ refers to men and $g = 2$ to women.

Based on previous research on body image issues in patients with burns, the researchers hypothesized, first, that the impact of burn severity on self-esteem would be higher in women with facial burns compared to men with facial burns, after controlling for symptoms of depression and anxiety (HADS) and age. More precisely, the researchers expected a negative relation between TBSA and self-esteem for both men and women, and they expected the effect to be stronger for women. Those expectations can be converted directly into inequality and order constraints, namely $\beta^1_{11} \leq 0,\ \beta^2_{11} \leq 0$ and $\beta^2_{11} \leq \beta^1_{11}$. Second, the researchers hypothesized a positive relation between TBSA and anxiety and depression symptoms for both men and women after controlling for rumination and age, and that the impact of TBSA on anxiety and depression symptoms would be higher in women. Therefore, the constraints for the second hypothesis are $\beta^1_{21} \geq 0,\ \beta^2_{21} \geq 0$ and $\beta^2_{21} \geq \beta^1_{21}$. Using Equations 5.4 and 5.5, these constraints can be written as:

$$\boldsymbol{L} = \begin{bmatrix} 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}, \tag{5.9}$$

$$\boldsymbol{\theta^a} = \begin{bmatrix} \beta^1_{21} & \beta^1_{11} & \beta^1_{13} & \beta^1_{21} & \beta^1_{22} & \beta^1_{23} & \beta^2_{21} & \beta^2_{11} & \beta^2_{13} & \gamma^2_{21} & \beta^2_{22} & \beta^2_{23} \end{bmatrix}^\top, \tag{5.10}$$

and $\boldsymbol{c} = \mathbf{0}$. Then by multiplying matrix $\boldsymbol{L}$ by vector $\boldsymbol{\theta^a}$ we can write the constrained hypothesis as:

$$H: \quad \begin{bmatrix} -\beta_{11}^1 \geq 0 \\ -\beta_{11}^2 \geq 0 \\ \beta_{11}^1 - \beta_{11}^2 \geq 0 \\ \beta_{21}^1 \geq 0 \\ \beta_{21}^2 \geq 0 \\ -\beta_{21}^1 + \beta_{21}^2 \geq 0 \end{bmatrix} = \begin{bmatrix} \beta_{11}^1 \leq 0 \\ \beta_{11}^2 \leq 0 \\ \beta_{11}^2 \leq \beta_{11}^1 \\ \beta_{21}^1 \geq 0 \\ \beta_{21}^2 \geq 0 \\ \beta_{21}^2 \geq \beta_{21}^1 \end{bmatrix} . \tag{5.11}$$

In the next section we will discuss a procedure for testing such a constrained hypothesis. It is important to note that, the `Informative-Testing()` function does not require to construct the complex $\boldsymbol{L}$ matrix in Equation 5.9 manually. After the next section, we show that the constraints can be specified by a user-friendly text-based description.

5.3 Procedure for testing constrained hypotheses in SEM

First, we start to discuss the parametric (Efron & Tibshirani, 1993, pp. 53–56) and the Bollen-Stine (Bollen & Stine, 1993, pp. 120–122) bootstrap approaches for obtaining a plug-in p value. Second, we introduce the genuine double bootstrap (Beran, 1988) method for adjusting the plug-in p value or alpha level.

5.3.1 Bootstrapping

An often used procedure for comparing the fit of nested models, for example H_{A0} versus H_{A1}, is the likelihood ratio (LR) statistic for hypothesis test Type A. This is defined as:

$$LR = -2log\left[\frac{L(\boldsymbol{\Sigma}(\boldsymbol{\theta}_{H_{A0}}) \mid \boldsymbol{Y})}{L(\boldsymbol{\Sigma}(\boldsymbol{\theta}_{H_{A1}}) \mid \boldsymbol{Y})}\right] , \tag{5.12}$$

where L is the likelihood probability of the observed data $\boldsymbol{Y}$ as a function of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and where $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the estimated model implied covariance matrix under H_{A0} and H_{A1}. For hypothesis test Type B, H_{A0} and H_{A1} are replaced by H_{B0} and H_{B1} respectively.

When the null and/or alternative hypothesis involves order/inequality constraints on the parameters, then the null-distribution of the LR statistic with multivariate normal data turns out the be $\bar{\chi}^2$-distributed (chi-square-bar) (Silvapulle & Sen, 2005). That is a weighted sum of chi-squared distributions where the weights can be estimated via Monte Carlo simulations [3] or via the procedure described in (Shapiro, 1988) when dealing with linear regression and (only) linear constraints. Alternatively, the p value of the statistic can be computed directly via bootstrapping (Van de Schoot et al., 2010), which is called a plug-in p value and is denoted by $\hat{p}$.

This can be done by two types of bootstrap methods, namely by parametric ($\hat{p}_{par}$) and Bollen-Stine ($\hat{p}_{bs}$) bootstrapping.[4] The plug-in p value usually refers to the parametric p value, however, for the sake of convenience we use the term plug-in p value also for the Bollen-Stine p value.

First, plug-in p value $\hat{p}_{par}$ can be obtained by parametric bootstrapping and can be summarized by the following steps for a hypothesis test of Type A:

1. Estimate $\boldsymbol{\theta}$ under the null-hypothesis H_{A0} using the observed data, resulting in $\boldsymbol{\Sigma}(\boldsymbol{\theta}_{HA0})$. Also, estimate $\boldsymbol{\theta}$ under H_{A1} resulting in $\boldsymbol{\Sigma}(\boldsymbol{\theta}_{HA1})$ and compute the LR value for the observed data as shown in Equation 5.12, which is denoted by LR^{obs}.

2. Draw $B_r^1 = B_1^1, \ldots, B_R^1$ bootstrap samples of size N from a known population distribution, say a multivariate normal distribution, using the estimated model implied covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta}_{HA0})$. Superscript 1 denotes the first-level bootstrap samples.

3. Estimate $\boldsymbol{\theta}_r$ for each bootstrap sample B_r^1 under H_{A0} and H_{A1}.

[3] Monte Carlo simulation is defined as a resampling technique that randomly generates samples from a known population distribution, such as the multivariate normal distribution (e.g., the parametric bootstrap). Non-parametric bootstrap procedures are similar to Monte Carlo simulations but the samples are drawn from the actual data and are therefore called resampling techniques (e.g., the Bollen-Stine bootstrap).

[4] A third procedure exists, called the naive, or simple, bootstrap, but as shown in (Bollen & Stine, 1993, pp. 117–119) it is inaccurate for testing the LR statistic for structural equation models, since the bootstrap sample should only reflect sampling variability and possibly non-normality, but not model misfit.

4. Compute for each B_r^1 sample the LR statistic. This results in a vector of R LR values, denoted by $LR_r^{boot} = LR_1^{boot}, \ldots, LR_R^{boot}$.

5. To calculate the plug-in p value compute

$$\hat{p} = \frac{\sum_{r=1}^{R} I(LR_r^{boot} > LR^{obs})}{R} , \tag{5.13}$$

where I is the indicator function equaling 1 if the expression inside the brackets is true and 0 otherwise.

Parametric bootstrapping is a powerful method when the underlying assumption of the population distribution is satisfied. For example, for continuous data following a multivariate normal distribution. If this assumption holds the parametric bootstrap approach is expected to have a better accuracy (Gentle, Härdle, & Mori, 2004, p. 469). When this assumption is violated then the Bollen-Stine bootstrap approach would lead to more accurate results, since no underlying population distribution is assumed. The Bollen-Stine method is simply a non-parametric bootstrap where data are transformed in accordance with the null-hypothesis. Consequently, any non-normality of the data is preserved and therefore also retained in each bootstrap sample.

For computing the Bollen-Stine plug-in p value $\hat{p}_{bs}$ only the first two steps are different compared to the parametric bootstrap:

1. Transform the observed data matrix so that its covariance structure is in accordance with the null-hypothesis.

2. Draw $B_r^1 = B_1^1, \ldots, B_R^1$ bootstrap samples of size N from the transformed data, and proceed with step 3 from the parametric bootstrap approach.

To transform the data, we can use

$$\boldsymbol{Z} = \boldsymbol{Y}\boldsymbol{S}^{-1/2}\boldsymbol{\Sigma}(\boldsymbol{\theta}_{HA0})^{1/2}, \tag{5.14}$$

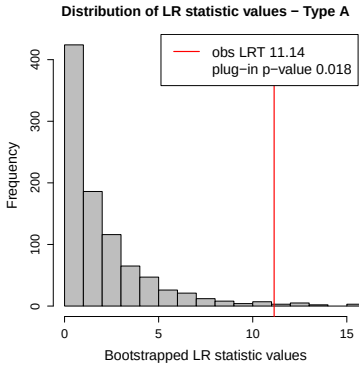
where $\boldsymbol{Z}$ is the transformed data, $\boldsymbol{Y}$ denotes the $N \times p$ data matrix of the centered observed variables, and $\boldsymbol{S}$ denotes the sample covariance matrix of $\boldsymbol{Y}$ (Bollen & Stine, 1993, pp. 120).

5.3.2 Facial burn example continued

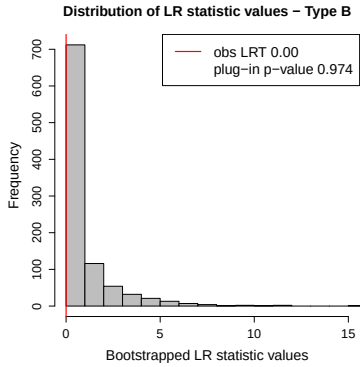
For the facial burns example Figure 5.2a and 5.2b display the result of the bootstrap for hypothesis test Type A and Type B. Note that the result applies for both bootstrap approaches. On the x-axis the LR values are given. Observe that most values are close to zero, since we sampled from the null-distribution. The solid vertical line represents the location of LR^{obs}. The plug-in p value is the proportion of LR^{boot} values on the right-hand side of the LR^{obs}. For a graphical representation of the parametric bootstrap see (Van de Schoot et al., 2010) and (Van de Schoot & Strohmeier, 2011). The two procedures previously described are repeated for hypothesis test Type B. However, estimating $\boldsymbol{\theta}$ under the constrained hypothesis H_{B0} is more complex than under the equally constrained hypothesis H_{A0}. Computationally, linear equality constraints are generally easier to deal with than linear inequality constraints. Linear equality constraints result in a dimension reduction of the parameter vector. The resulting unconstrained problem can be solved using simpler methods for unconstrained optimization (Nocedal & Wright, 2006, Ch. 17).

5.3.3 Double bootstrapping

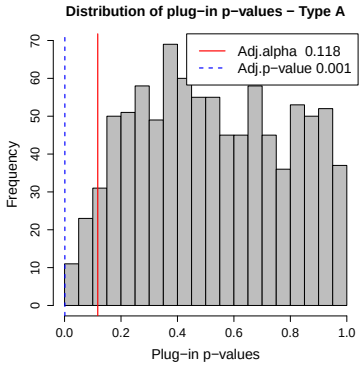
In the previous section we introduced the plug-in p value based on parametric and Bollen-Stine bootstrapping. A well-known property of the p value, and also of our plug-in p value, is that it is asymptotically a uniform distribution [0,1] under the H_0, such that $P(p < \alpha \mid H_{A0}) = \alpha$. This is also true for $\hat{p}$ when $R \to \infty$. However, when constraints are imposed on $\boldsymbol{\theta^a}$, it appears that $P(\hat{p} < \alpha \mid H_{A0}) \neq \alpha$. The parametric as well as the non-parametric bootstrap are not consistent if a parameter is on a boundary of the parameter space defined by (non)linear inequality constraints or a mixture between (non)linear inequality and equality constraints (Andrews, 2000). If this is the case either α needs to be adjusted or $\hat{p}$ needs to be adjusted. Here, we discuss the genuine double bootstrap approach to adjust α and $\hat{p}$. We show how to obtain an adjusted alpha level, denoted by α^* and an adjusted plug-in p value, denoted by $\hat{p}^*$. In cases where it is necessary to make a distinction between the parametric and Bollen-Stine bootstrap we add the subscripts "par" and "bs" to $\hat{p}^*$ or α^*.



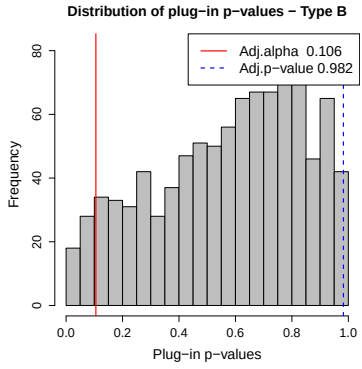
(a) Result of bootstrapping the LR values for Type A for the facial burns example.



(b) Result of bootstrapping the LR values for Type B for the facial burns example.



(c) Result of the genuine double bootstrap for Type A for the facial burns example.



(d) Result of the genuine double bootstrap for Type B for the facial burns example.

Figure 5.2: (a) Result of bootstrapping the LR values for the facial burns example for hypothesis test Type A. The solid line represents the LR^{obs} value. The proportion of LR_r^{boot} values on the right-hand side of the solid line is the plug-in p value $\hat{p}$. (b) See (a), but now for hypothesis test Type B. (c) Result of the genuine double bootstrap for the facial burns example for hypothesis test Type A. The non-uniform distribution of plug-in p values means that adjustment of α or $\hat{p}$ is necessary. The solid line represents the adjusted alpha level α^* and the dashed line the adjusted plug-in p value $\hat{p}^*$. (d) See (c), but now for hypothesis test Type B.

For the genuine double bootstrap the following steps are needed for obtaining α^* or $\hat{p}^*$:

1. Draw $B_r^1 = B_1^1, \ldots, B_R^1$ bootstrap samples of size N using either the parametric or Bollen-Stine bootstrap. Compute $\hat{p}_{par}$ or $\hat{p}_{bs}$ as defined in Equation 5.13.

2. Each of the B_r^1 is treated as an observed data set from which second-level parametric or Bollen-Stine bootstrap samples $B_{rs}^2 = B_{r1}^2, \ldots, B_{rS}^2$ are drawn [5].

3. Use the second-level B_{rs}^2 bootstrap samples to compute R plug-in p values resulting in a vector of $\hat{p}_r = \hat{p}_1, \ldots, \hat{p}_R$.

The result, so far, is a vector of R plug-in p values. The distribution of these plug-in p values should be uniform when $R \to \infty$. If this is the case then adjusting α or $\hat{p}$ is not necessary. For the facial burns example the distributions for hypothesis tests Type A and Type B in Figure 5.2c and 5.2d are clearly not uniform and adjustment is necessary. For test Type A, $P(\hat{p} < .12 \mid H_{A0}) \neq .05$ and for test Type B, $P(\hat{p} < .11 \mid H_{B0}) \neq .05$. Now we continue with computing α^* or $\hat{p}^*$:

4a. The adjusted alpha level α^* is calculated by first ordering $\hat{p}_r$ from small to large followed by computing the x^{th} percentile, which is typically the 5^{th} percentile at a significance level of 5%. In Figure 5.2c and 5.2d the solid lines show the adjusted alpha levels.

4b. According to (Nankervis, 2005), the adjusted p value $\hat{p}^*$ is calculated by:

$$\hat{p}^* = \frac{\sum_{r=1}^{R} I(\hat{p}_r < \hat{p})}{R} \ . \tag{5.15}$$

In Figure 5.2c and 5.2d the dashed lines show the adjusted plug-in p values.

[5] The choice of S is always a tradeoff between precision and practical use. If we use as many as 1000 samples for both R and S, then we would need as many as 10^6 samples. (Davison & Hinkley, 2008, Ch. 5.6) suggest that $S = 249$ would be safe.

For a graphical representation of the genuine double bootstrap see (Van de Schoot et al., 2010). The steps previously described are repeated for hypothesis test Type B.

In the next section, we will discuss the `InformativeTesting()` function in more detail.

5.4 An overview of the `InformativeTesting` function in `R` package `lavaan`

At the time of writing, the `InformativeTesting()` function is included in `lavaan`, a free and open source `R` package for *la*tent *va*riable *an*alysis (Rosseel, 2012b) (`http://lavaan.org`).

Before we can test the constrained hypothesis of our facial burns example as defined in Equation 5.11 we need to go through 5 easy steps:

Step 1 is to call the `lavaan` library:

```
R> library("lavaan")
```

Step 2 is to load the observed data into `R`. The data can be a data frame containing the observed variables or a sample covariance full matrix with an optional mean vector. For our example the data are loaded into `R` as follows:

```
R> FacialBurns <- read.csv("burns.csv")
```

Step 3 is to convert the theoretical model in Figure 5.1 into `lavaan` syntax. The input model is specified by a text-based description called the `lavaan` model syntax and includes the overall model without constraints.

```
R> burnsModel <- ' Selfesteem ~ Age + c(m1, f1)*TBSA + HADS
                          HADS ~ Age + c(m2, f2)*TBSA + RUM '
```

where `m1, f1, m2` and `f2` are arbitrary labels which are necessary for imposing the constraints.

Step 4 is to convert the constraints into `lavaan` syntax. For the sake of convenience we do not use the Greek letter β with subscripts and superscript, but simple labels. Therefore let $\beta^1_{11} =$ `m1`, $\beta^2_{11} =$ `f1`, $\beta^1_{21} =$ `m2` and

β_{21}^2 = `f2`. The constraints are specified by a text-based description, called the `lavaan` constraints syntax and describe the linear order/inequality constraints imposed on the model. A major advantage of this text-based description is that users do not have to specify the complex $\boldsymbol{L}$ matrix (see 5.9) themselves. Then, the constraints are defined as follows:

```
R> burnsConstraints <- ' m1 < 0
                         f1 < 0
                         f1 < m1
                         m2 > 0
                         f2 > 0
                         f2 > m2 '
```

Note that these constraints equal to the right hand side of equation 5.11. Also note that, it is only necessary to specify the overall model and the constraints, the equality constrained model H_{A0} and the unconstrained model H_{B1} are generated automatically by the `InformativeTesting()` function and thus need not to be specified. For more information about how to create the model and constraints syntax, see the `lavaan` manual (Rosseel, 2012b).

Step 5 is to set up the necessary `InformativeTesting()` function arguments. For an overview of all function arguments see `?InformativeTesting`. The first argument to `InformativeTesting()` is the model defined in step 3. The second argument is the observed data. The third argument is the constraints imposed on the model in step 4. The fourth and fifth arguments define the number of bootstrap draws and the number of double bootstrap draws respectively. In our example we used $R = 1000$ and $S = 249$. The group argument specifies the grouping variable, which is in our case the variable name `"Sex"` in the data frame. The parallel and ncpus arguments are needed to use parallel processing. In this example we used 8 cores for the computations.

```
R> burnsIT <- InformativeTesting(burnsModel, data = FacialBurns,
                                 constraints = burnsConstraints,
                                 R = 1000, double.bootstrap.R = 249,
                                 group = "Sex",
                                 parallel = "multicore", ncpus = 8)
```

The `InformativeTesting()` function bootstraps LR values and it returns an object of class `"InformativeTesting"` for which a `plot()`

method is available, which is discussed later. By default the `InformativeTesting()` function uses the Bollen-Stine bootstrap approach (`type = "bollen.stine"`) and the genuine double bootstrap for adjusting the plug-in p value (`double.bootstrap = "standard"`). However, users can easily switch to the parametric bootstrap (`type = "parametric"`) or turn the double bootstrap off (`double.bootstrap = "no"`). Furthermore, by default the `InformativeTesting()` function generates $R = 1000$ bootstrap draws and returns a vector with the bootstrapped LR values (`return.LRT = TRUE`). For the genuine double bootstrap `double.bootstrap.R` $= 249$ double bootstrap samples are drawn. Note that for the `"standard"` double bootstrap by default S $= 249$ and a significance level of 5% is used to compute α^* (`double.bootstrap.alpha = 0.05`).

In the next section we will discuss the `print()` and `plot()` methods for the `InformativeTesting()` function using the facial burns example.

5.4.1 Facial burn example continued: `print()` and `plot()`

Perhaps the most informative method to view the results is `plot()`. The `plot()` function plots the distributions of the bootstrapped LR values and also the distributions of the plug-in p values in case of the genuine double bootstrap. The `plot()` method can be called without additional arguments to plot all available plots.

Separate plots for the distribution of LR values or the plug-in p values can be requested. For the distribution of LR values the argument `type = "lr"` is added (see Figure 5.2a and 5.2b) and for the distribution of plug-in p values the argument `type = "ppv"` is added, see Figure 5.2c and 5.2d. The first argument to `plot()` is the returned object from the `InformativeTesting()` function.

```
R> plot(burnsIT)
```

For the `plot()` results for the facial burn example see Figure 5.2a, 5.2b, 5.2c and 5.2d. The default plot arguments can be overruled by the user to adjust the plots. For example, the axes labels, main title, number of breaks, and colors can be adjusted. For all available options see `?plot.InformativeTesting`.

A table of the `print()` function is displayed with the results of the facial burn example.

```
R> burnsIT

InformativeTesting: Order/Inequality Constrained Hypothesis Testing:

  Variable names in model         : Selfesteem HADS Age TBSA RUM
  Number of variables             : 5
  Number of groups                : 2
  Used sample size per group      : 60 17
  Used sample size                : 77
  Total sample size               : 118

  Estimator                       : ML
  Missing data                    : listwise
  Bootstrap method                : bollen.stine
  Double bootstrap method         : standard


  Type A test: H0: all restriktions active (=)
          vs. H1: at least one restriktion strictly true (>)
         Test statistic: 11.1374, adjusted p-value: 0.0011 (alpha = 0.05)
                                unadjusted p-value: 0.0182 (alpha = 0.1176)

  Type B test: H0: all restriktions true
          vs. H1: at least one restriktion false
         Test statistic: 0.0000, adjusted p-value: 0.9824 (alpha = 0.05)
                               unadjusted p-value: 0.9742 (alpha = 0.1055)
```

The results for hypothesis test Type A, see also Figures 5.2a and 5.2c, show that the equality constrained hypothesis H_{A0} is rejected ($LR^{obs} = 11.1374,\ \hat{p}^* = .001,\ \alpha = .05$). In other words, the observed LR statistic $LR^{obs} = 11.1374$, see solid line Figure 5.2a, is more extreme than we would expect by chance. The adjusted plug-in p value $\hat{p}^*$ is the proportion of plug-in p values on the right-hand side of the dashed line. The results for Type B, see also Figures 5.2b and 5.2d, show that the constrained hypothesis H_{B0} cannot be rejected ($LR^{obs} = 0,\ \hat{p}^* = .982,\ \alpha = .05$). The observed LR statistic, $LR^{obs} = 0$, indicates that the constrained hypothesis does not fit significantly worse than the unconstrained. More precisely, a LR value of 0 indicates that no constraints are violated. Therefore, we can conclude that the results show strong evidence for the constrained hypothesis stated in Equation 5.11. In other words, there exists a negative relation between TBSA and self-esteem for both men and women, and the relation is stronger for women. It is also confirmed that the relation between TBSA and HADS is positive for both men and women and that the relation is stronger for women. A visual inspection of the constrained

or unconstrained model parameters reinforce our conclusion. This can be done as follows for the unconstrained model. Some parts of the output are removed due to its length.

```
R> summary(burnsIT$fit.B1)

Group 1 [1]:

Regressions:
                   Estimate
  Selfesteem ~
    Age                0.019
    TBSA      (m1)    -0.148
    HADS               -0.475
  HADS ~
    Age                0.070
    TBSA      (m2)     0.143
    RUM                1.036


Group 2 [2]:

Regressions:
                   Estimate
  Selfesteem ~
    Age                0.106
    TBSA      (f1)    -0.241
    HADS               -0.453
  HADS ~
    Age               -0.010
    TBSA      (f2)     0.254
    RUM               -0.656
```

The constrained parameter estimates can be requested as follows:

```
R> summary(burnsIT$fit.A1)
```

A tutorial function with the R-code and the data from the facial burns example is available online at [MASKED]. For more information and a gentle introduction to informative hypotheses we recommend the book of (Hoijtink, 2012), and also the papers from (Van de Schoot, Hoijtink, &

Romeijn, 2011; Van de Schoot & Strohmeier, 2011) and (Van de Schoot & Wong, 2011) for more applied examples.

5.5 Concluding remarks

In classical null-hypothesis testing researchers can only find indirect evidence for their specific hypotheses if the null-hypothesis is rejected. Therefore, we believe that evaluating informative hypotheses, by means of imposing order/inequality constraints on the parameters of a statistical model, allow researchers to directly evaluate their expectations and get more insightful results compared to testing the classical null-hypothesis against catch-all rivals. In addition, (Vanbrabant, Van de Schoot, & Rosseel, 2015) (and the references therein) have shown that substantial power can be gained when an increasing number of order/inequality constraints is included into the hypothesis. Researchers who are dealing with small samples in particular may benefit from this power gain.

Hypothesis test Type A can reject the null-hypothesis H_{A0} even if the alternative hypothesis H_{A1} is violated by the data. Rejecting H_{A0} does not mean that H_{A1} is true. Note that, this applies also to classical null-hypothesis testing. The power of hypothesis tests Type A and Type B is centered in the alternative hypothesis H_1. It is only under H_0 that their type I errors is close to the nominal level. In spite of this, hypothesis test Type A can be useful since hypothesis test Type B cannot make a distinction between equality and inequality constraints. Hypothesis test Type B plays a crucial role in constraint misspecification.

The `InformativeTesting()` function discussed in this paper is the first software routine for testing order/inequality constrained hypotheses in SEM. If researchers want to test a constrained hypothesis, then the procedure with the `InformativeTesting()` function for `lavaan` is easier to use and faster compared to the procedure proposed in (Van de Schoot et al., 2010).

However, testing constrained hypotheses in SEM is due to the genuine double bootstrap procedure computationally very expensive. Therefore, computational time remains a limitation and procedures to decrease it further are investigated.

Furthermore, we conducted a small simulation study in which we in-

vestigated the performance of the Bollen-Stine and parametric bootstrap approaches in terms of type I errors ($\alpha = .05$). We have set up a simulation design in which we varied the sample size and the normality of the data. We chose a small sample size of $N = 50$ and a large sample size of $N = 500$. We generated normal and very non-normal data with a skewness of 1.50 and a kurtosis of 3.75. The results show that the Bollen-Stine bootstrap outperforms the parametric bootstrap in case of non-normal data. We recommend to use the parametric bootstrap only in case of normal distributed samples. Currently, the parametric bootstrap for the `InformativeTesting()` function is only valid for continuous data following a multivariate normal distribution.

Finally, we advise to use the genuine double bootstrap and only to switch the double bootstrap off when exploration is the goal of the analysis.

Acknowledgments

The first author is a PhD fellow of the research foundation Flanders (FWO) at Ghent university (Belgium) and at Utrecht University (The Netherlands). The second author is supported by a grant from the Netherlands organization for scientific research: NWO-VENI-451-11-008. Data for the empirical sample were obtained for a study funded by the Dutch Burn Foundation (Grant no 05.109). We thank all participating patients and the research team in the respective burn centers (Groningen: M. Bremer, G. Bakker; Beverwijk: A. Boekelaar; Rotterdam: A. van de Steenoven, H. Hofland).

References

Andrews, D. (2000). Inconsistency of the boostrap when a parameter is on the boundary of the parameter space. *Econometrica*, *68*, 399–405.

APA. (2000). *Diagnostic and Statistical Manual of Mental Disorders* (4^{th} ed.). Washington, DC: American Psychiatric Association.

Barlow, R., Bartholomew, D., Bremner, H., & Brunk, H. (1972). *Statistical inference under order restrictions*. New York: Wiley.

Beran, R. (1988). Prepivoting test statistics: A bootstrap view of asymptotic refinements. *Journal of the American Statistical Association*, *83*, 687–697. doi: doi:10.1080/01621459.1988.10478649

Bollen, K. A., & Stine, R. (1993). *Bootstrapping Goodness-of-Fit Measures in Structural Equation Models* (K. A. Bollen & J. Scott Long, Eds.). Sage Publications, Newbury Park.

Davison, A., & Hinkley, D. (2008). *Bootstrap metods and their application* (10th ed.). Cambridge University press.

Efron, E., & Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman & Hall.

Gentle, J., Härdle, W., & Mori, Y. (Eds.). (2004). *Handbook of computational statistics: Concepts and methods*. Spring-Verlag: Berlin.

Hoijtink, H. (2012). *Informative Hypotheses: Theory and Practice for Behavioral and Social Scientists*. Boca Raton, FL: Taylor & Francis.

Hoogewerf, C. J., van Baar, M. E., Middelkoop, E., & van Loey, N. E. (2014). Impact of facial burns: relationship between depressive symptoms, self-esteem and scar severity. *General Hospital Psychiatry*, *36*(3), 271–276.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality Constrained Analysis Of Variance: A Bayesian Approach. *Psychological Methods*, *10*, 477–493. doi: doi:10.1037/1082-989X.10.4.477

Kuiper, R., Klugkist, I., & Hoijtink, H. (2010). A Fortran 90 Program for Confirmatory Analysis of Variance. *Journal of Statistical Software*, *34*, 1–31.

Mulder, J., Hoijtink, H., & de Leeuw, C. (2012). BIEMS: A Fortran 90 Program for Calculating Bayes Factors for Inequality and Equality Constrained Models. *Journal of Statistical Software*, *46*, 1–38.

Muthen, L., & Muthen, B. (2010). *Mplus: Statistical analysis with latent variables: User's guide*. Los Angeles.

Nankervis, J. (2005). *Stopping Rules for Double Bootstrap Tests.* (Working paper, University of Essex)

Nocedal, J., & Wright, S. (2006). *Numerical Optimization* (2nd ed.;

V. Mikosh, S. Resnick, & S. Robinson, Eds.). Spring-Verlag: New York.

Nolen-Hoeksema, S., Wisco, B. E., & Lyubomirsky, S. (2008). Rethinking Rumination. *Perspectives on Psychological Science*, *3*, 400–24. doi: doi:10.1111/j.1745-6924.2008.00088.x

R Development Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. (ISBN 3-900051-07-0)

Robertson, T., Wright, F. T., & Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. New York: Wiley.

Rosenberg, M. (1965). *Society and the Adolescent Self-image*. Princeton, NJ: Princeton University Press.

Rosseel, Y. (2012a). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. Retrieved from `http://www.jstatsoft.org/v48/i02/`

Rosseel, Y. (2012b). lavaan: An R Package for Structural Equation Modeling and More [Computer software manual]. Department of Data Analysis Ghent University (Belgium). Retrieved from `http://lavaan.org`

Shapiro, A. (1988). Towards a unified theory of inequality constrained testing in multivariate analysis. *International Statistical Review*, *56*, 49–62. doi: doi:10.2307/1403361

Silvapulle, M., & Sen, P. (2005). *Constrained statistical inference: Order, inequality, and shape restrictions*. Hoboken, NJ: Wiley.

Spinhoven, P., Ormel, J., Sloekers, P. P. A., Kempen, G., Speckens, A. E. M., & VanHemert, A. M. (1997). A validation study of the Hospital Anxiety and Depression Scale (HADS) in different groups of Dutch subjects. *Psychological Medicine*, *27*, 363–370. doi: doi:10.1017/S0033291796004382

Strahan, E. J., Wilson, A. E., Cressman, K. E., & Buote, V. M. (2006). Comparing to Perfection: How Cultural Norms for Appearance Affect Social Comparisons and Self-image. *Body image*, *3*, 211–27. doi: doi:10.1016/j.bodyim.2006.07.004

Thombs, B. D., Notes, L. D., Lawrence, J. W., Magyar-Russell, G., Bresnick, M. G., & Fauerbach, J. A. (2008). From Survival to Socialization: A Longitudinal Study of Body Image in Survivors of Severe Burn Injury. *Journal of Psychosomatic Research*, *64*, 205–12. doi: doi:10.1016/j.jpsychores.2007.09.003

Van de Schoot, R., Hoijtink, H., & Deković, M. (2010). Testing inequality constrained hypotheses in SEM models. *Structural Equation Mod-*

eling, *17*, 443–463. doi: doi:10.1080/10705511.2010.489010

Van de Schoot, R., Hoijtink, H., Hallquist, M., & Boelen, P. (2012). Bayesian evaluation of inequality-constrained hypotheses in SEM models using mplus. *Structural Equation Modeling*, *19*, 593–609. doi: doi:10.1080/10705511.2012.713267

Van de Schoot, R., Hoijtink, H., Mulder, J., Van Aken, M. A. G., Orobio de Castro, B., Meeus, W., & Romeijn, J. (2011). Evaluating Expectations about Negative Emotional States of Aggressive Boys using Bayesian Model Selection. *Developmental Psychology*, *47*, 203–212. doi: doi:10.1037/a0020957

Van de Schoot, R., Hoijtink, H., & Romeijn, J. (2011). Moving beyond traditional null hypothesis testing: evaluating expectations directly. *Frontiers in Psychology*, 1–5. doi: doi:10.3389/fpsyg.2011.00024

Van de Schoot, R., & Strohmeier, D. (2011). Testing informative hypotheses in SEM increases power: An illustration contrasting classical hypothesis testing with a parametric bootstrap approach. *International Journal of Behavioral Development*, *35*, 180–190. doi: doi:10.1177/0165025410397432

Van de Schoot, R., & Wong, T. (2011). Do antisocial young adults have a high or a low level of self-concept? *Self and Identity, First published on: 24 January 2011 (iFirst) doi:10.1080/15298868.2010.517713*.

Vanbrabant, L., Van de Schoot, R., & Rosseel, Y. (2015). Constrained statistical inference: sample-size tables for anova and regression. *Frontiers in Psychology*, *5*, 1–8. doi: doi:10.3389/fpsyg.2014.01565

Wiechman, S. A., Ptacek, J., Patterson, D. R., Gibran, N. S., Engrav, L. E., & Heimbach, D. M. (2001). Rates, Trends, and Severity of Depression after Burn Injuries. *Journal of Burn Care & Rehabilitation*, *22*, 417–24. doi: doi:10.1097/00004630-200111000-00012

6

An introduction to restriktor: informative hypothesis testing for AN(C)OVA and linear models

Many researchers have specific expectations about the relation between the means of different groups or between (standardized) regression coefficients. For example, in an experimental setting, the comparison of two or more treatment groups may be subject to order constraints (e.g., $H_1 : \mu_1 < \mu_2 < \mu_3 = \mu_4$). In practice, hypothesis H_1 is usually tested using a classical one-way ANOVA with additional pairwise comparisons if the corresponding F-test is significant. In this tutorial paper, we introduce the freely available `R` package `restriktor` for evaluating order-constrained hypothesis directly. The procedure is illustrated by seven examples.

6.1 Introduction

In almost all psychological fields, researchers have specific expectations about the relation between the means of different groups or between (stan-

dardized) regression coefficients. In experimental psychology, it is often tested whether the mean reaction time *increases* or *decreases* for different treatment groups (see e.g., Kofler et al., 2013). In clinical trials, it is often tested whether a particular treatment is *better* or *worse* than other treatments (see e.g., Roberts, Roberts, Jones, & Bisson, 2015). In observational studies, researchers often have clear ideas about whether the direction of the effects are *positive* or *negative* (see e.g., Richardson & Abraham, 2012). Testing such specific expectations directly is known under various names, such as one-sided testing, constrained statistical inference, isotonic regression, and informative hypothesis testing. For the remainder of this paper, we will refer to this kind of analysis as informative hypothesis testing (IHT, Hoijtink, 2012).

Many applied researchers are already familiar with IHT in the context of the classical one-sided t-test, where one mean is restricted to be greater or smaller than a fixed value (e.g., $\mu_1 > 0$) or another mean (e.g., $\mu_1 < \mu_2$). This readily extends to the AN(C)OVA and multiple regression (e.g., linear, logistic, Poisson) setting where more than one constraint can be imposed on the (adjusted) means or regression coefficients (Silvapulle & Sen, 2005).

IHT has several benefits compared to classical null-hypothesis significance testing. First, testing specific expectations directly does not require multiple significance tests (Hoijtink, 2012; Klugkist, van Wesel, & Bullens, 2011; Van de Schoot et al., 2011). In this way, we avoid an inflated type I error or a decrease in power when a significance level α correction is used. Second, to avoid multiple testing issues with ordered means, an ANOVA is often combined with contrasts to directly test the specific pattern. However, contrast tests are not the same as informative hypothesis tests (Baayen, Klugkist, & Mechsner, 2012). Third, incorporating order constraints in the analysis will result in substantially more power (e.g., Bartholomew, 1961a, 1961b; Kuiper & Hoijtink, 2010; Perlman, 1969; Robertson, Wright, & Dykstra, 1988; Van de Schoot & Strohmeier, 2011; Vanbrabant, Van de Schoot, & Rosseel, 2015). Vanbrabant et al. (2015) showed for ordered means and multiple one-sided regression coefficients that a sample-size reduction up to 50% can be gained.

Hypothesis testing in the linear model and in the AN(C)OVA model assumes that the residuals are normally and independently distributed. Although the well-known F-test statistic, which is often used in linear re-

gression and AN(C)OVA is size robust (close to their nominal significance level α) for deviations from the normality assumption, it can have substantial consequences for the power (Schrader & Hettmansperger, 1980; Silvapulle, 1992; Wilcox, 2016). The reason for the lower power is that non-normal error distributions are more likely to contain extreme observations (e.g., outliers) and these outliers can increase the sample residual variance estimate (i.e., the scale) substantially. Even when the deviations are small enough to go undetected by distribution normality checks (Rutherford, 2001, Chp. 9). Robust hypothesis testing is a powerful alternative. Robustness is achieved by down-weighting extreme observations to have less influence on the estimates (Huber, 1981; Huber & Ronchetti, 2009; Maronna, Martin, & Yohai, 2006). MM-estimation (Yohai, 1987), is perhaps the most frequently applied robust regression technique today and it is widely available in statistical software (e.g., SAS, Stata, various `R` packages).

Several software routines are available for testing informative hypotheses in the frequentist framework. Ordered means may be evaluated by the software routine 'Confirmatory ANOVA' (Kuiper, Klugkist, & Hoijtink, 2010). An extension for linear regression models is available in the `R` (R Development Core Team, 2016) package `ic.infer` (Grömping, 2010). Order constraints may also be evaluated by the statistical software SAS/STAT® (SAS Institute Inc, 2008) using the PLM procedure. Model selection under order constraints can be performed using the software routine 'GORIC' (Kuiper, Hoijtink, & Silvapulle, 2012). However, these procedures are rather complex, since the constraint matrix must almost always be constructed manually. In addition, the procedures are limited to ordered means or the standard linear regression model. In this current paper we introduce the open-source and freely availabe `R` package `restriktor` (`http://restriktor.org`). We will show that `restriktor` is easy to use and more flexible than the existing procedures.

In the remainder of this article, we demonstrate for seven examples how to evaluate informative hypotheses using `restriktor`. For each example, we show (1) how to set up the constraint syntax, (2) how to test the informative hypothesis, and (3) how to interpret the results. Many of the `restriktor` options are discussed gradually over the various examples. In the first example, we impose order constraints on the means of a one-way ANOVA model. In the second example, we reanalyze the

first example but using robust methods to deal with outliers. In the third example, we impose order constraints on the means of an ANOVA model, where we take a small effect-size into account. In the fourth example, we impose order contraints on the adjusted means of an ANCOVA model. In the fifth example, we impose order constraints on the standardized regression coefficients of a linear model. In the sixth example, we impose order constraints on three covariate-conditional effects of gender on the outcome variable. In the last example, we demonstrate how to evaluate the informative hypothesis H_1 using model selection. Instead of comparing H_1 only against the unconstrained hypothesis H_{u}, we will also include competing informative hypotheses. The corresponding models are evaluated based on their fit and complexity using the generalized order-restricted information criterion (GORIC). After the examples, we discuss some additional options of `restriktor`. To ensure applicability of this paper, the datasets for each of the examples are available in the `restriktor` package.

6.2 Example 1. order-constrained one-way ANOVA

Consider the data in Table 6.1. These data denote a persons' decrease in aggression level between week 1 (intake) and week 8 (end of training) for four different treatment groups of anger management training, namely (1) no training, (2) physical training, (3) behavioral therapy, and (4) a combination of physical exercise and behavioral therapy. The purpose of the study was to test the assumption that the exercises would be associated with a reduction in the mean aggression levels. In particular, the hypothesis of interest was $H_1 : \mu_{\text{No}} < \{\mu_{\text{Physical}} = \mu_{\text{Behavioral}}\} < \mu_{\text{Both}}$. This hypothesis states that the decrease in aggression levels is smallest for the "no training" group, larger for the "physical training" and "behavioral therapy" group, with no preference for either method, and largest in the "combination of physical exercise and behavioral therapy" group (Hoijtink, 2012, p. 5–6).

In practice, hypothesis H_1 is usually evaluated with an ANOVA, where the null-hypothesis $H_0 : \mu_{\text{No}} = \mu_{\text{Physical}} = \mu_{\text{Behavioral}} = \mu_{\text{Both}}$ is tested against the unconstrained-hypothesis H_{u} : not all four means are equal. The results from the global F-test revealed that the four means are not

equal ($F_{(4,36)} = 18.62$, $p < .001$). At this point, we do not know anything about the ordering of the means. Therefore, the next step would be to use pairwise comparisons with corrections for multiple testing (Westfall, Tobias, & Wolfinger, 2011, e.g., Bonferroni, FDR, Tukey). The results with FDR (False Discovery Rate) adjusted p-values showed three significant ($p \leq .05$) mean differences (MD), namely between the 'Behavioral-No' exercises (MD = 3.3, $p = .001$), the 'Behavioral-Physical' exercises (MD = 2.3, $p = .018$) and the 'Both-Physical' exercises (MD = 3.3, $p = .001$). A graphical representation of the means is shown in Figure 6.1 (see filled circles). Based on the results of the global F-test and the pairwise comparisons, it would not be an easy task to derive an unequivocal conclusion about hypothesis H_1.

In what follows, we show all steps and the `restriktor` syntax to evaluate the informative hypothesis H_1 directly. Before we continue, we need to install the `R` package `restriktor`. To install `restriktor`, start up `R`, and type:

```
install.packages("restriktor")
```

If the `restriktor` package is installed, the package needs to be loaded into `R`. This can be done by typing:

```
library(restriktor)
```

If the package is loaded, the following startup message should be displayed:

```
## This is restriktor 0.1-80.711
## restriktor is BETA software!  Please report any bugs.
```

A more detailed description about how to get started with `restriktor` can be found online at `http://restriktor.org/gettingstarted.html`.

Step 1. set up the constraint syntax

In `R`, categorical predictors are represented by 'factors'. For example, the 'Group' variable has four factor levels: 'No', 'Physical', 'Behavioral' and 'Both'. In addition, the factor levels are presented in alphabetical order

and it may therefore be convenient to re-order the levels. This can be done in R by typing:

```
AngerManagement$Group <- factor(AngerManagement$Group,
                                levels = c("No","Physical",
                                           "Behavioral",
                                           "Both"))
```

In restriktor there are two ways to construct the constraint syntax. First, and probably also the easiest way is to use the factor-level names preceded by the factor name (e.g., GroupNo). Order constraints are defined by means of inequality constraints ($<$, or $>$) or by equality constraints ($==$). The constraint syntax is enclosed within single quotes. Then, for hypothesis H_1 the constraint syntax might looks as follows:

```
myConstraints1 <- ' GroupNo         < GroupPhysical
                    GroupPhysical  == GroupBehavioral
                    GroupBehavioral < GroupBoth '
```

A second method is to construct the constraint matrix manually. The corresponding restriktor code might look as follows:

```
myConstraints1 <- rbind(c( 0, 1, -1, 0),
                        c(-1, 1,  0, 0),
                        c( 0, 0, -1, 1))
```

Note that the first row should be treated as an equality contraint. This can be done in the restriktor() function by setting the neq = 1 argument. We will not further elaborate on this method because it is error prone to inexperienced users. For the interested reader we refer to the restriktor website or to the restriktor() function help file, which can be found in R by typing ?restriktor.

Step 2. test the informative hypothesis

The iht() function is used for informative hypothesis testing. The minimal requirements for this function are a constraint syntax (e.g., see myConstraints1) and a fitted unconstrained model. Currently, iht() can

deal with the standard linear model (`lm`), the robust linear model (`rlm`) and the generalized linear model (`glm`). Since, an AN(C)OVA model is a special case of the multiple regression model we can use the linear model for our ANOVA example. Then, we can fit the unconstrained linear model as follows:

```
fit_ANOVA <- lm(Anger ~ -1 + Group, data = AngerManagement)
```

The tilde $\sim$ is the regression operator. On the left-hand side of the operator we have the response variable `Anger` and on the right-hand side we have the factor `Group`. We removed the intercept (`-1`) from the model so that the estimates reflect the group means. The `AngerManagement` dataset is build-in the `restriktor` package and can be called directly. Information about importing your own dataset into `R` can be found online at `http://restriktor.org/tutorial/importdata.html`.

Next, we can test the informative hypothesis using the `iht()` function. This is done as follows:

```
iht(fit_ANOVA, constraints = myConstraints1)
```

The first argument to `iht()` is the fitted unconstrained linear model. The second argument is the constraint syntax `myConstraints1`. By default, the function prints an overview of all available hypothesis tests. The results are shown below.

```
Restriktor: restricted hypothesis tests ( 36 residual degrees of freedom ):

Multiple R-squared reduced from 0.674 to 0.608

Constraint matrix:
   GroupNo GroupPhysical GroupBehavioral GroupBoth    op rhs active
1:       0             1              -1         0    ==   0    yes
2:      -1             1               0         0    >=   0     no
3:       0             0              -1         1    >=   0     no


Overview of all available hypothesis tests:

Global test: H0: all parameters are restricted to be equal (==)
         vs. HA: at least one inequality restriction is strictly true (>)
       Test statistic: 25.4061,   p-value: <0.0001
```

```
Type A test: H0: all restrictions are equalities (==)
         vs. HA: at least one inequality restriction is strictly true (>)
       Test statistic: 25.4061,   p-value: <0.0001

Type B test: H0: all restrictions hold in the population
         vs. HA: at least one restriction is violated
       Test statistic: 7.2687,   p-value: 0.04518

Note: All tests are based on a mixture of F-distributions
      (Type C test is not applicable because of equality restrictions)
```

At the top of the output the constraint-matrix is shown. This matrix is constructed internally based on the text-based constraint syntax but could of course have been constructed manually. The 'active' column indicates if a constraint is violated or not. If no constraints are active, this would mean that all constraints are in line with the data. Next, an overview of the available hypothesis tests is given. By default `restriktor` uses the $\bar{F}$ (F-bar) test-statistic (Kudô, 1963; Wolak, 1987). The $\bar{F}$-statistic is an adapted version of the classical F-statistic and can deal with order constraints. To ensure readability of this paper, its technical details are discussed in Appendix M.1. The global hypothesis test is comparable with the classical global/omnibus test, where *all* parameters but the intercept equal zero under the null-hypothesis and it is tested against the order-constrained hypothesis. Under the null of hypothesis test Type A, only the parameters that are involved in the order-constrained hypothesis (here all) are constrained to be equal and it is tested against the order-constrained hypothesis. For hypothesis test Type B, the null-hypothesis is the order-constrained hypothesis and it is tested against the unconstrained hypothesis, although some equality constraints (if present) may be preserved under the alternative hypothesis. Rejecting the null-hypothesis would mean that at least one order constraint is violated. A more detailed output for each hypothesis test, such as the estimates under both hypotheses can be obtained by adding the argument `type = "Global"`, `"A"` or `"B"` to the `iht()` function. Note that there exists another hypothesis test called Type C (not applicable here because of an equality constraint). This test is based on the union-intersection principle. Its power is generally poor in case of a relatively large number of constraints (Grömping, 2010). Hypothesis test Type C is added to complete the set of tests but we will not further discuss it. For the interested reader we

refer to Silvapulle & Sen, 2005, Chp. 5.3.

Step 3. interpret the results

To evaluate the informative hypothesis H_1, we first conduct hypothesis test Type B. Not rejecting this hypothesis test would mean that the order constraints are in line with the data. The results from hypothesis test Type B, however, show that hypothesis H_1 is rejected in favor of the best fitting hypothesis ($\bar{F}^{B}_{(0,1,2;36)} = 7.27$, p = .045) [1]. In other words, the constraints are not supported by the data and we conclude that the informative hypothesis H_1 does not hold.

Estimation and inference of the restricted estimates

Instead of testing the informative hypothesis, the restricted estimates might be of interest. In this case, the `restriktor()` function can be used:

```
restr_ANOVA <- restriktor(fit_ANOVA,
                          constraints = myConstraints1)
```

The first argument to `restriktor()` is the fitted unconstrained linear model `fit_ANOVA`. The second argument is the constraint syntax `myConstraints1`. By default, the `print()` function prints a brief overview of the restricted estimates:

```
print(restr_ANOVA)

Call:
conLM.lm(object = fit_ANOVA, constraints = myConstraints1)

restriktor (0.1-80.711): restricted linear model:

Coefficients:
       GroupNo    GroupPhysical  GroupBehavioral        GroupBoth
         -0.20             1.95             1.95             4.10
```

[1] The null-distribution is a mixture of F-distributions mixed over the degrees of freedom. Therefore, in this example, the p-value $\Pr(\bar{F} \geq \bar{F}_{obs})$ approximately equals $w_0 \Pr(F_{0,36} \geq \bar{F}_{obs}) + w_1 \Pr(F_{1,36} \geq \bar{F}_{obs}/1) + w_2 \Pr(F_{2,36} \geq \bar{F}_{obs}/2)$, where $\Pr(F_{0,36} \geq \bar{F}_{obs})$ equals 0 by definition. Hence the notation $\bar{F}_{(0,1,2;36)}$. For more information on how to compute the mixing weights w_i see Appendix M.1.

We can clearly see that, the `GroupPhysical` and the `GroupBehavioral` estimates are constrained to be equal. If desired, a more extensive output can be requested. This is done as follows:

```
summary(restr_ANOVA)

Call:
conLM.lm(object = fit_ANOVA, constraints = myConstraints1)

Restriktor: restricted linear model:

Residuals:
   Min     1Q Median     3Q    Max
-3.100 -1.275 -0.025  1.200  5.050

Coefficients:
                Estimate Std. Error t value  Pr(>|t|)
GroupNo         -0.20000    0.65233 -0.3066 0.7609210
GroupPhysical    1.95000    0.46127  4.2275 0.0001544 ***
GroupBehavioral  1.95000    0.46127  4.2275 0.0001544 ***
GroupBoth        4.10000    0.65233  6.2851 2.895e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.0629 on 36 degrees of freedom
Standard errors: standard
Multiple R-squared reduced from 0.674 to 0.608

Generalized Order-Restricted Information Criterion:
  Loglik  Penalty    Goric
-84.1621   2.8918 174.1079
```

The output shows the restricted estimates (here the group means) and the corresponding standard errors, t-test statistics and two-sided p-values. The output also shows information about the type of computed standard errors. In this case, conventional standard errors are computed but heteroskedastic robust standard errors are also available. The multiple $R^2 = .674$ refers to the unconstrained model and the $R^2 = .608$ refers to the order-constrained model. Both are equal, only if all constraints are in line with the data. The last part of the output provides information for model selection. This will be discussed in example 7.

6.2.1 Example 2. order-constrained robust one-way ANOVA

The results in the previous example were obtained under the ANOVA assumptions that the residuals were normally and independently distributed. In this example, we rerun the ANOVA example using robust IHT. We show that ignoring non-normality lead to spurious conclusions. The steps to run the anger management training example with robust MM-estimators are identical to the ANOVA example, except for fitting the unconstrained model. Instead of the standard linear model, we now use the robust linear model. The unconstrained robust linear model needs to be fitted using the `rlm()` (W. N. Venables and B. D. Ripley, 2002) function in `R`. This can be done as follows:

```
fit_rANOVA <- rlm(Anger ~ -1 + Group, data = AngerManagement,
                  method = "MM")
```

Note that by default the `rlm` function uses M-estimation. It is easy to switch to MM-estimation by adding the `method = "MM"` argument.

Then, evaluating hypothesis H_1 using robust IHT can be done as follows:

```
iht(fit_rANOVA, constraints = myConstraints1)
```

The output of the `iht()` function is shown in Appendix N.1. In this case, `restriktor` uses by default the robust $\bar{F}_{mm}$ test-statistic (Silvapulle, 1992). Its technical details are discussed in Appendix M.2. The results show that hypothesis H_1 is now *not* rejected in favor of the unconstrained hypothesis ($\bar{F}^{B}_{mm(1,2,3;36)} = 6.49$, p = .062). This is an illustration that ignoring non-normality may result in spurious conclusions regarding the direction of the effects. If hypothesis test Type B is not rejected, a second hypothesis test is needed. The reason is that hypothesis test Type B cannot make a distinction between inequality and equality constraints. In the statistical literature, this hypothesis test is often called Type A, where hypothesis H_0 is tested against the order-constrained hypothesis H_1. The results from hypothesis test Type A show that hypothesis H_0 is rejected in favor of the order-constrained hypothesis H_1 ($\bar{F}^{A}_{mm(0,1,2;36)} = 21.55$, p < .001). If we combine the results of robust hypothesis test

Type B and robust hypothesis test Type A, we can conclude that we have found evidence in favor of the informative hypothesis H_1 [2].

6.2.2 Example 3. Ordered-constrained means with effect-sizes

The p-value is not a good measure for the size of an effect (Nickerson, 2000). Therefore, in an AN(C)OVA the question should actually be whether the differences between the group means are relevant? To answer this question, the popular effect-size measure Cohen's d (Cohen, 1988) can be used and is given by: $d = (\mu_{\max} - \mu_{\min})/\sigma_\epsilon$, where $\mu_{\max}$ is the largest of the k means and $\mu_{\min}$ is the smallest of the k means, and σ_ϵ is the pooled standard deviation within the populations. According to Cohen, values of 0.2, 0.5 and 0.8 indicate a small, medium and large effect, respectively.

In this example, we use the Zelazo, Zelazo, and Kolb (1972) dataset. The data consist of ages in months at which a child starts to walk for four treatment groups. For simplicity we only consider three treatment groups. The excluded group is the 'Control' group. The first treatment group (Active) received a special walking exercise for 12 minutes per day beginning at age 1 week and lasting 7 weeks. The second group (Passive) received daily exercises but not the special walking exercises. The third group (No) were checked weakly for progress but they did not receive any special exercises. The purpose of the study was to test the claim that the walking exercises are associated with a reduction in the mean age at which children start to walk.

If we ignore the effect-sizes, the informative hypothesis can be formulated as: $H_2 : \mu_{\text{Active}} < \mu_{\text{Passive}} < \mu_{\text{No}}$. The results from hypothesis test Type B ($\bar{F}^{B}_{(0,1,2;14)} = 0$, p = 1) and hypothesis test Type A ($\bar{F}^{A}_{(0,1,2;14)} = 5.978$, p = .028) provide evidence in favor of the informative hypothesis. However, for a practical implementation of the treatments the mean differences between the groups should at least indicate a small effect. To answer this question, we reformulate hypothesis H_2 such that the effect-

[2] We are aware that strictly speaking, null-hypothesis significance testing never provides 'evidence' for the null and that the results provide indirect evidence for the informative hypothesis but this is not the place to discuss null-hypothesis significance testing subtleties.

sizes are included. The pooled within group standard deviation equals 1.516:

$$H_2^d = \begin{array}{lll} (\mu_{\text{Passive}} & - \mu_{\text{Active}}) & / \ 1.516 > 0.2 \\ (\mu_{\text{No}} & - \mu_{\text{Passive}}) & / \ 1.516 > 0.2. \end{array}$$

This hypothesis states that we expect at least 0.2×1.516 standard deviations between the means, which indicates a small effect-size. Next, we show how to evalute this informative hypothesis.

Step 1. set up the constraint syntax

Again, we use the factor-level names preceded by the factor name to construct the constraint syntax. The effect-sizes can be easily computed within the constraint syntax using the arithmetic operator `/`:

```
myConstraints2 <- ' (GroupPassive - GroupActive ) / 1.516 > 0.2
                    (GroupNo      - GroupPassive) / 1.516 > 0.2 '
```

Step 2. test the informative hypothesis

The original dataset consists of four treatment groups. Since, we excluded the 'Control' group, we need to take a subset of the original data. The `subset()` function in `R` is the easiest way to select observations. This can be done in `R` by typing:

```
subData <- subset(ZelazoKolb1972, Group != "Control")
```

The first argument to `subset()` is the original dataset. The second argument excludes the observations from the 'Control' group using the `!=` (not equal to) operator. Then, the unconstrained linear model can be fit as follows:

```
fit_ANOVAd <- lm(Age ~ -1 + Group, data = subData)
```

Next, we test the informative hypothesis using the fitted unconstrained model `fit_ANOVAd` and the constraint syntax `myConstraints2`:

```
iht(fit_ANOVAd, constraints = myConstraints2)
```

The output of the `iht()` function can be found in Appendix N.2.

Step 3. interpret the results

The results from hypothesis test Type B ($\bar{F}^{B}_{(0,1,2;14)} = 0$, p = 1) and hypothesis test Type A ($\bar{F}^{A}_{(0,1,2;14)} = 3.19$, p = .089) show that if we include a small effect-size in the informative hypothesis, the initial significant results become irrelevant. This clearly demonstrates the importance of including effect-sizes in the hypothesis.

6.2.3 Example 4. Order-constrained adjusted means - ANCOVA

The anger management training example discussed in example 1 also included a covariate; it was not considered in the introduction for simplicity. The covariate provides information about a persons' age (ranging from 18 to 27). The full `AngerManagement` dataset is displayed in Table 6.2. In contrast to ANOVA, where informative hypotheses are formulated in terms of group means, informative hypotheses in an ANCOVA are formulated in terms of adjusted means to account for differences between the groups with respect to one or more covariates. Thus, if we take the covariate 'age' into account the informative hypothesis can be formulated as $H_1^{\text{adj}} : \mu^{\text{adj}}_{\text{No}} < \{\mu^{\text{adj}}_{\text{Physical}} = \mu^{\text{adj}}_{\text{Behavioral}}\} < \mu^{\text{adj}}_{\text{Both}}$ (μ_j^{adj} denotes the population adjusted mean in group j). A graphical representation of the covariate adjusted means is shown in Figure 6.1 (see unfilled circles).

Step 1. set up the constraint syntax

For hypothesis H_1^{adj} the constraint syntax is identical to the constraint syntax for the ANOVA example and looked as follows:

```
myConstraints1 <- ' GroupNo         < GroupPhysical
                    GroupPhysical  == GroupBehavioral
                    GroupBehavioral < GroupBoth '
```

Step 2. test the informative hypothesis

Before we fit the unconstrained model, we center the covariate 'Age' at its average to obtain adjusted mean [3] estimates. This is done in R by typing:

```
AngerManagement$Age_Z <- AngerManagement$Age -
                                    mean(AngerManagement$Age)
```

Then, we can fit the unconstrained linear model as follows:

```
fit_ANCOVA <- lm(Anger ~ -1 + Group + Age_Z,
                 data = AngerManagement)
```

Next, we can test the informative hypothesis using the `iht()` function. This is done as follows:

```
iht(fit_ANCOVA, constraints = myConstraints1)
```

The results are shown in Appendix N.3.

Step 3. interpret the results

As a reminder, in order to find evidence for the informative hypothesis H_1^{adj}, we do not want to reject hypothesis test Type B. The results, however, show that hypothesis test Type B is rejected in favor of the unconstrained hypothesis ($\bar{\text{F}}^{\text{B}}_{(1,2,3;35)} = 8.50$, p = .028). Therefore, we can conclude that the imposed order constraints are not supported by the data. The results from the robust hypothesis test Type B lead to the same conclusion ($\bar{\text{F}}^{\text{B}}_{\text{mm}(1,2,3;35)} = 7.70$, p = .037).

6.2.4 Example 5. Order-constrained (standardized) linear regression coefficients

In this example, we show how order constraints can be imposed on the standardized regression coefficients of a linear model. We use the `Exam`

[3] The general formula to compute the adjusted means is: $Y_{\text{adj}j} = \bar{Y}_j - \beta(\bar{\text{Z}}_j - \bar{\text{Z}}_G)$, where $\bar{Y}_j$ are the unadjusted means in group j, $\bar{Z}_j$ are the covariate means in group j, $\bar{Z}_G$ is the general covariate mean, and β is the common within-group regression coefficient.

dataset displayed in Table 6.3. The model relates students' 'exam scores' (Scores, with a range of 38 to 82) to the 'averaged point score' (APS, with a range of 18 to 28), the amount of 'study hours' (Hours, with a range of 25 to 61), and 'anxiety score' (Anxiety, with a range of 13 to 91). It is hypothesized that APS is the strongest predictor, followed by 'study hours' and 'anxiety scores', respectively. In symbols, this informative hypothesis can be written as $H_3 : \beta_{\text{APS}} > \beta_{\text{Hours}} > \beta_{\text{Anxiety}}$ (β denotes the standardized regression coefficient). Since, the hypothesis is in terms of which predictor is stronger, we should be aware that the predictor variables are measured on a different scale. Using the unstandardized coefficients might lead to spurious conclusions. Therefore, the predictor variables should be standardized [4] first. This can be done in `R` by typing:

```
Exam$Hours_Z   <- (Exam$Hours   - mean(Exam$Hours))   / sd(Exam$Hours)
Exam$Anxiety_Z <- (Exam$Anxiety - mean(Exam$Anxiety)) / sd(Exam$Anxiety)
Exam$APS_Z     <- (Exam$APS     - mean(Exam$APS))     / sd(Exam$APS)
```

Step 1. set up the constraint syntax

We can refer to covariates simply by their name (e.g., `APS_Z`). Then, the constraint syntax corresponding H_3 might look as follows:

```
myConstraints3 <- ' APS_Z   > Hours_Z
                    Hours_Z > Anxiety_Z '
```

Step 2. test the informative hypothesis

Next, we fit the unconstrained linear model. The response variable is 'Scores' and the predictor variables are the three centered covariates:

```
fit_exam <- lm(Scores ~ APS_Z + Hours_Z + Anxiety_Z,
               data = Exam)
```

The informative hypothesis H_3 can be evaluated using the unconstrained model `fit_exam` and the constraint syntax `myConstraints3`:

[4] Standardized regression coefficients can be obtained by standardizing all the predictor variables before including them in the model. For example: $\text{Z}(\text{APS}_i) = \big(\text{APS}_i$ - mean(APS)$\big)$ / sd(APS), where sd is the standard deviation.

```
iht(fit_exam, constraints = myConstraints3)
```

The output is shown in Appendix N.4.

Step 3. interpret the results

The results from hypothesis test Type B show that the order-constrained hypothesis is not rejected in favor of the unconstrained hypothesis ($\bar{F}^{B}_{(0,1,2;16)} = 0$, p = 1). The results from hypothesis test Type A show that the null-hypothesis is rejected in favor of the order-constrained hypothesis ($\bar{F}^{A}_{(0,1,2;16)} = 12.38$, p = .003). Thus, we have found strong evidence in favor of the informative hypothesis H_3.

6.2.5 Example 6. Testing for order-constrained effects

Here, we show how order constraints can be imposed between newly defined parameters. The original data are based on two cohort studies in children from 0 to 4 and 8 to 18 years old with burns and their parents (e.g., Bakker, Van der Heijden, Van Son, & Van Loey, 2013; Egberts et al., 2016). Since, the original data are not publicly accessible, we simulated data from the orginal model parameters. This simulated dataset is available in `restriktor`. For illustrative reasons we focus only on the data provided by the mother. The final sample consists of mothers of 278 children. Boys represent 68.7% of the sample. The response variable is parental post-traumatic stress symptoms (PTSS) and was measured with the Impact of Event Scale (Horowitz, Wilner, & Alvarez, 1979). Moreover, for the current illustration we included five predictor variables in the dataset: a child's gender (0 = boys, 1 = girls) and age, the estimated percentage total body surface area affected by second or third degree burns (i.e., TBSA, with a range of 1-72% in the current sample) and parental guilt [0-4] and anger [0-4] feelings in relation to the burn event. The model relates PTSS to the five predictor variables and can be written as a linear

function:

$$\begin{aligned}\text{PTSS} \sim \beta_0 + \beta_1\text{gender} + \beta_2\text{age} + \beta_3\text{guilt} + \beta_4\text{anger} + \beta_5\text{TBSA} \\ + \beta_6\text{gender} \times \text{guilt} \\ + \beta_7\text{gender} \times \text{anger} \\ + \beta_8\text{gender} \times \text{TBSA}\end{aligned}$$

where β_0 is the intercept, β_1 to β_5 are the regression coefficients for the main-effects and β_6 to β_8 are the regression coefficients for the interaction-effects.

We hypothesized that the gender-effect would increase for simultaneously higher levels of guilt, anger and TBSA. To test this informative hypothesis, we selected three different settings for guilt, anger and TBSA, namely a small, a medium and a large level. For illustrative reasons, we chose for the small level the values 0, 0, 1 for guilt, anger and TBSA respectively. For the medium level we chose their mean values which are 2.02, 2.06, and 8.35, respectively, and for the large level we chose 4, 4, and 20, respectively. Then, the resulting three effects (small, medium, large) can be calculated as follows respectively:

$$\begin{aligned}\text{smallEffect} &= \beta_1 + \beta_6 0 + \beta_7 0 + \beta_8 1 \\ \text{mediumEffect} &= \beta_1 + \beta_6 2.02 + \beta_7 2.06 + \beta_8 8.35 \\ \text{largeEffect} &= \beta_1 + \beta_6 4 + \beta_7 4 + \beta_8 20.\end{aligned}$$

Note that each effect reflects a mean difference between boys and girls. Then, the informative hypothesis can be expressed as:

$$H_4 : \text{smallEffect} < \text{mediumEffect} < \text{largeEffect}.$$

Step 1. set up the constraints syntax

A convenient feature of the `restriktor` constraint syntax is the option to define new parameters, which take on values that are an arbitrary function of the original model parameters. This can be done using the `:=` operator. In this way, we can compute the desired effects and impose order constraints among these effects. Then, the constraint syntax might look as follows:

```
myConstraints4 <- ' ## define the effects
                    smallEffect  := gender + 0*gender.guilt +
                                    0*gender.anger +
                                    1*gender.TBSA

                    mediumEffect := gender + 2.02*gender.guilt +
                                    2.06*gender.anger +
                                    8.35*gender.TBSA

                    largeEffect  := gender + 4*gender.guilt +
                                    4*gender.anger +
                                   20*gender.TBSA

                    ## impose the order constraints
                    smallEffect  < mediumEffect
                    mediumEffect < largeEffect '
```

It is important to note that variable/factor names of the interaction effects in objects of class `lm` and `rlm` contain a semi-colon (`:`) between the variable names (e.g., `gender:guilt`). To use these parameters in the constraint syntax, the semi-colon must be replaced by a dot (`.`) (e.g., `gender.guilt`).

Step 2. test the informative hypothesis

Based on outlier diagnostics [5] we identified 13 outliers (approximately 4.7% of the data). Therefore, we use robust methods. The unconstrained robust linear model using MM-estimation can be fitted as follows:

```
fit_rburns <- rlm(PTSS ~ gender*guilt + gender*anger +
                  gender*TBSA + age,
                  data = Burns, method = "MM")
```

On the right-hand side of the regression operator ($\sim$) we included the three interaction-effect using the `*` operator. The main-effects are in this way automatically included. Note that the interaction operator `*` is not an arithmetic operator as used in the constraint syntax. Then, the informative hypothesis can be evaluated as follows:

[5] The outliers were identified with robust Mahalanobis distances larger than the 99.5% quantile of a χ^2_8 distribution.

```
iht(fit_rburns, constraints = myConstraints4)
```

The output can be seen in Appendix N.5.

Step 3. interpret the results

The results from hypothesis test Type B ($\bar{F}^{B}_{MM(0,1,2;269)} = 0$, p = 1) show that the order-constrained hypothesis is not rejected in favor of the unconstrained hypothesis. The results from hypothesis test Type A show that the null-hypothesis is rejected in favor of the order-constrained hypothesis ($\bar{F}^{A}_{MM(0,1,2;269)} = 5.35$, p = .044). Hence, we can conclude that the data provide enough evidence that the gender-effect increases for higher levels of guilt, anger and TBSA.

Noteworthy, the non-robust results from hypothesis test Type A would have let to a different conclusion, namely that the null-hypothesis would not have been rejected in favor of the order-constrained hypothesis ($\bar{F}^{A}_{(0,1,2;269)} = 3.65$, p = .107). Again, this clearly demonstrates that ignoring outliers may result in misleading conclusions.

6.2.6 Example 7. Model selection under order constraints

In the previous examples, we used hypothesis testing to evaluate the informative hypotheses. In this example, we demonstrate the generalized order-restricted information criterion (GORIC), which is a modification of the Akaike information criterion (AIC, (Akaike, 1998). The GORIC can be used to evaluate competing hypotheses based on their fit (i.e., likelihood) and complexity (i.e., (in)equality constraints). The complexity provides information about the simplicity of the model. The unconstrained model is the most complex model, where no prior information about the parameters (e.g., means, regression coefficients, variance) is known. The model with equality constraints is on the other hand the simplest model and the model with order constraints lies somewhere in between.

Reconsider the order-constrained hypothesis $H_1 : \mu_{\text{No}} < \{\mu_{\text{Physical}} = \mu_{\text{Behavioral}}\} < \mu_{\text{Both}}$ from example 1. To test this informative hypothesis, we evaluated it against the competing unconstrained hypothesis

(hypothesis test Type B). However, instead of using the unconstrained-hypothesis as competing hypothesis, it is also possible to specify other order-constrained hypotheses. The GORIC can be used to evaluate a set of informative hypotheses. Suppose, we want to evaluate the following set of informative hypotheses:

$$
\begin{aligned}
H_0 &: \mu_{\text{No}} = \mu_{\text{Physical}} = \mu_{\text{Behavioral}} = \mu_{\text{Both}} \\
H_1 &: \mu_{\text{No}} < \{\mu_{\text{Physical}} = \mu_{\text{Behavioral}}\} < \mu_{\text{Both}} \\
H_2 &: \mu_{\text{No}} < \mu_{\text{Physical}} < \mu_{\text{Behavioral}} < \mu_{\text{Both}} \\
H_{\text{u}} &: \mu_{\text{No}} \ , \ \mu_{\text{Physical}} \ , \ \mu_{\text{Behavioral}} \ , \ \mu_{\text{Both}}.
\end{aligned}
$$

Note that it is recommended to also include the unconstrained hypothesis H_{u} in the set to avoid choosing a weak/bad model. The model with the lowest GORIC value is the preferred one. To improve the interpretation, we also compute the GORIC weights, which are comparable to the Akaike weights and reflect the support for each model in the set.

Step 1. set up the constraint syntaxes

First, we construct the syntax for each hypothesis, except for the unconstrained hypothesis of course:

```
myConstraints1 <- ' GroupNo         == GroupPhysical
                    GroupPhysical   == GroupBehavioral
                    GroupBehavioral == GroupBoth '

myConstraints2 <- ' GroupNo         < GroupPhysical
                    GroupPhysical  == GroupBehavioral
                    GroupBehavioral < GroupBoth '

myConstraints3 <- ' GroupNo         < GroupPhysical
                    GroupPhysical   < GroupBehavioral
                    GroupBehavioral < GroupBoth '
```

Step 2. compute the GORIC values and GORIC weights

First, we fit the unconstrained model. The model is identical to the one discussed in example 1 and was specified as follows:

```
fit_ANOVA <- lm(Anger ~ -1 + Group, data = AngerManagement)
```

Second, we fit all three restricted models and the unconstrained model using the `restriktor()` function:

```
restr1_ANOVA <- restriktor(fit_ANOVA, constraints = myConstraints1)
restr2_ANOVA <- restriktor(fit_ANOVA, constraints = myConstraints2)
restr3_ANOVA <- restriktor(fit_ANOVA, constraints = myConstraints3)
restr4_ANOVA <- restriktor(fit_ANOVA)
```

Finally, we use the `goric()` function to compute the log-likelihood, penalty (complexity), GORIC value, and the GORIC weight for each model. The input for the `goric()` function are the four fitted `restriktor` objects:

```
goric(restr1_ANOVA, restr2_ANOVA, restr3_ANOVA, restr4_ANOVA)
```

The function prints a table with all the results:

```
          model   loglik  penalty   goric  goric_weights
1  restr1_ANOVA  -93.401   2.0000  190.80   0.0000061596
2  restr2_ANOVA  -84.162   2.8918  174.11   0.0259843803
3  restr3_ANOVA  -80.484   3.0833  167.13   0.8491068095
4  restr4_ANOVA  -80.484   5.0000  170.97   0.1249026505
```

Step 3. interpret the results

The first column, shows the name of the model. The second column, shows the log-likelihood for each model. Note that the log-likelihood cannot make a distinction between model H_2 (`restr3_ANOVA`, fully ordered) and model H_{u} (`restr4_ANOVA`, unconstrained). The third column, shows the complexity, where the unconstrained-model has the highest penalty term (5) and the equality-constrained model the lowest (2). To clarify, the penalty is computed as follows: for the unconstrained model four means and one variance needs to be estimated, and for the equality constrained model only one mean and one variance need to be estimated. The penalty for models with inequality constraints is a bit more difficult to compute but it depends on the mixing (chi-bar-square) weights

(Kuiper, 2011). The fourth column, shows the GORIC values. The model with the lowest GORIC value is the preferred one. Note that the GORIC value for the unconstrained model renders to the AIC. The last column, shows the GORIC weights and reflect the support of each model in the set. If we want to compare model H_1 (`restr2_ANOVA`) with model H_2 (`restr3_ANOVA`) we can examine the ratio of the two corresponding GORIC weights: $0.849/0.026 = 32.654$. This means that model H_2 is about 32.654 times more likely than model H_1. In addition, model H_2 is about 6.792 (0.849/0.125) more likely than the unconstrained model H_{u}. Hence, we can concluded that model H_2 (`restr3_ANOVA`) is the preferred one.

6.3 `restriktor` options

All results in this paper were obtained by the default settings of the software package `restriktor`. In many scenarios they work well but if desired they can readily be adjusted. Instead of conventional standard errors, heteroskedastic robust Huber-White (Huber, 1967; White, 1980) standard errors or refinements of this can be computed by adding the argument `se = "HC"` (refinements: `"HC1", "HC2", "HC3", "HC4", "HC4m", "HC5"`) to the `restriktor()` function. Also, bootstrapped (standard or model-based) standard errors can be requested. The hypothesis tests were evaluated using the $\bar{\text{F}}$ test-statistic for the linear model and the $\bar{\text{F}}_{\text{mm}}$ test statistic for the robust linear model. Currently, `restriktor` can also compute a likelihood ratio test-statistic and a score test-statistic. They can be computed by adding the argument `test = "LRT"` or `"score"` to the `iht()` function. Nevertheless, preliminary simulation results show that the $\bar{\text{F}}$ and the $\bar{\text{F}}_{\text{mm}}$ test-statistics perform best in terms of size and power, even in small samples. In this paper, we only discussed models where the dependent variable is continuous. However, `restriktor` also supports models where the dependent variable is dichotomous or ordinal. In this case, the unconstrained model needs to be fitted using the `glm()` (generalized linear model) function in `R`. For all available options of the `restriktor()` function and the `iht()` function we refer to the `restriktor` website or to the help file in `R` by typing `?restriktor` or `?iht`, respectively.

It has to be noted that the restriktor package is not finished yet. But it is already very useful for most users. The package is actively maintained and new options are being added. We advise to monitor the restriktor website (http://restriktor.org) in order to be up-to-date.

6.4 Discussion

For more than a century, classical null-hypothesis testing has dominated the social and behavioral sciences. Nevertheless, this statistical approach has been heavily criticized in the psychological literature (Cohen, 1994; Cumming, 2008; Nickerson, 2000; Wagenmakers, 2007). Several of these critiques focus on the argument that the classical null-hypothesis does not provide the behavioral and social researcher with the needed information they want. Therefore, in this paper, we discussed (robust) informative hypothesis testing as a powerful alternative for evaluating expectations that cannot be expressed by the classical null-hypothesis (e.g., $H : \mu_1 < \mu_2 < \mu_3 = \mu_4$). For seven examples, we showed how informative hypotheses could be evaluated using the `R` package `restriktor`.

We only discussed frequentist methods for evaluating informative hypotheses. Of course, all the examples could have been perfectly evaluated in the Bayesian framework (Berger & Mortera, 1999; Gu, Mulder, Deković, & Hoijtink, 2014; Hoijtink, 2012; Klugkist, Laudy, & Hoijtink, 2005; Mulder, Hoijtink, & Klugkist, 2010) but we believe that the frequentist methods are a welcome addition to the applied user's toolbox and may help convince applied users to include order constraints in their hypothesis. The reason is that evaluating informative hypotheses using Bayesian statistics might be too big a step for researchers who are unfamiliar with both methods. In addition, robust informative hypothesis testing as discussed in this paper does not seem to exist in the Bayesian framework (yet).

In conclusion, informative hypothesis testing has shown to have major benefits compared to classical null-hypothesis testing. Unfortunately, applied researchers have been unable to use these methods because user-friendly freely available software and a clear tutorial were not available. As we have shown in this paper, these tools are ready to be used.

Acknowledgments

The first author is a PhD fellow of the research foundation Flanders (FWO) at Ghent university (Belgium) and at Utrecht University (The Netherlands). The second author is supported by a grant from the Netherlands organization for scientific research: NWO: VIDI-452-14-006.

Table 6.1: Persons' decrease in aggression levels for four treatment groups.

Group 1 Nothing	Group 2 Physical	Group 3 Behavioral	Group 4 Both
1	1	4	7
0	0	7	2
0	0	1	3
1	2	4	1
-1	0	-1	6
-2	1	2	3
2	-1	5	7
-3	2	0	3
1	2	3	5
-1	1	6	4

Note: these data originated from Hoijtink, 2012.

Table 6.2: Persons' decrease in aggression levels for four treatment groups and covariate age.

Group 1 Nothing		Group 2 Physical		Group 3 Behavioral		Group 4 Both	
Anger	Age	Anger	Age	Anger	Age	Anger	Age
1	18	1	23	4	21	7	21
0	20	0	24	7	22	2	22
0	21	0	19	1	23	3	23
1	22	2	20	4	25	1	25
-1	23	0	21	-1	26	6	24
-2	24	1	18	2	27	3	23
2	19	-1	20	5	23	7	26
-3	21	2	22	0	21	3	27
1	20	2	23	3	22	5	24
-1	22	1	21	6	25	4	23

Note: these data originated from Hoijtink, 2012.

Table 6.3: Exam scores, the amount of study hours, anxiety scores and the average point score (APS) for a group of 20 students.

Score	Hours	Anxiety	APS
62	40	40	24
58	31	65	20
52	35	34	22
55	26	91	22
75	51	46	28
82	48	52	28
38	25	48	18
55	37	61	20
48	30	34	18
68	44	74	26
62	32	54	24
62	40	61	24
72	61	26	26
58	35	13	24
65	45	54	20
42	30	58	20
68	39	62	24
68	47	39	26
58	41	57	22
72	46	17	28

Note: these data originated from `http://staff.bath.ac.uk/pssiw/stats2/examrevision.sav`

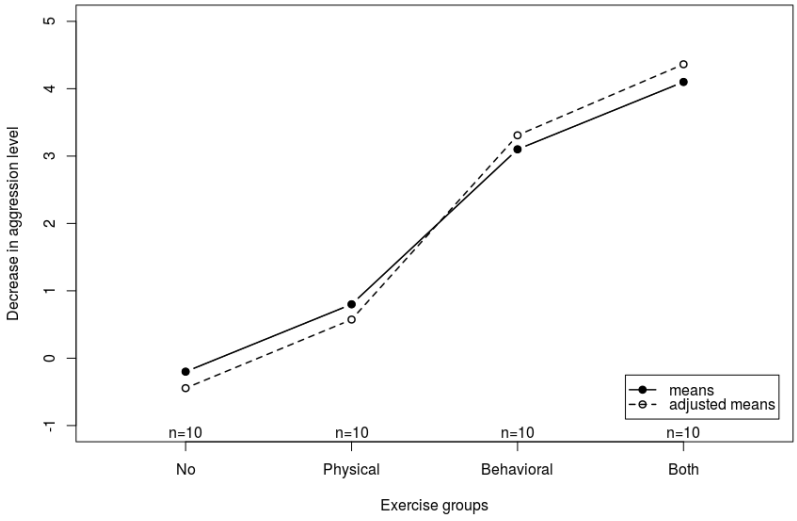


Figure 6.1: Means plot: reduction of aggression levels after 8 weeks of anger management training. The filled circles are the unadjusted means and the unfilled circles are the covariate 'age' adjusted means.

M
Test-statistics

Here, we discuss the non-robust $\bar{\mathrm{F}}$ test-statistic and the robust $\bar{\mathrm{F}}_{\mathrm{mm}}$ test-statistic. Moreover, we also discuss how to compute the p value. But first, we describe the linear regression model:

$$y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_{p-1} X_{i,p-1} + \epsilon_i,\ i = 1, \ldots, n. \quad \text{(M.1)}$$

We may express this in the familiar matrix form $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{X}$ is the design matrix, $\boldsymbol{\beta}$ is the vector with the regression coefficients $(\beta_0, \beta_1, \ldots, \beta_{p-1})$ and the vector $\boldsymbol{\epsilon}$ contains the random errors $(\epsilon_1, \ldots, \epsilon_n)$. Then, let $\hat{\boldsymbol{\beta}}$ be a vector with the unconstrained estimates, $\bar{\boldsymbol{\beta}}$ a vector with the estimates under the null model with equality constraints, and $\tilde{\boldsymbol{\beta}}$ a vector with the estimates of the inequality constrained optimization problem. To make a distinction between MM- and OLS-estimates, we added the subscript OLS. The symbol ^{T} denotes the transpose of a vector or matrix.

M.1 $\bar{\mathrm{F}}$ test-statistic

The $\bar{\mathrm{F}}$ test-statistic for hypothesis test Type A is given by

$$\bar{\mathrm{F}}^{\mathrm{A}} = (\bar{\boldsymbol{\beta}}_{\mathrm{ols}} - \tilde{\boldsymbol{\beta}}_{\mathrm{ols}})^T(\mathbf{X}^T\mathbf{X})(\tilde{\boldsymbol{\beta}}_{\mathrm{ols}} - \bar{\boldsymbol{\beta}}_{\mathrm{ols}})/\hat{\sigma}^2_{\mathrm{ols}}, \tag{M.2}$$

where $\hat{\sigma}^2_{\mathrm{ols}} = \sum_i^n (y_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_{\mathrm{ols}})^2/(n-p)$ is the error variance. For hypothesis test Type B the $\bar{\mathrm{F}}$ statistic is given by

$$\bar{\mathrm{F}}^{\mathrm{B}} = (\tilde{\boldsymbol{\beta}}_{\mathrm{ols}} - \hat{\boldsymbol{\beta}}_{\mathrm{ols}})^T(\mathbf{X}^T\mathbf{X})(\tilde{\boldsymbol{\beta}}_{\mathrm{ols}} - \hat{\boldsymbol{\beta}}_{\mathrm{ols}})/\hat{\sigma}^2_{\mathrm{ols}}. \tag{M.3}$$

M.2 $\bar{\mathrm{F}}_{\mathbf{mm}}$ test-statistic

MM-estimators are based on two loss functions ρ_1 and ρ_2 which determine the breakdown point (BDP) and the efficiency of the estimator respectively. Simply put, the BDP of a parameter estimate $\hat{\beta}_j$ is the largest proportion of irregularities that the data may contain such that $\hat{\beta}_j$ still gives some information about β_j (Maronna et al., 2006). Thus the higher the BDP the more robust the estimator. Theoretically, MM-estimators have a BDP of 50%. Let $\psi_j(\cdot) = \rho'_j(\cdot)$ for $j = 1, 2$ where the prime denotes differentiation. For both loss functions we use a Tukey biweight function which yields an MM-estimator that is robust to both outliers and (bad)-leverage points. To clarify, outliers are defined as extreme observations in the response space and bad-leverage points are defined as extreme observations in both the response *and* predictor space. The weights for Tukey's biweights are

$$\rho(\,e;c\,) = \left\{ \begin{array}{ll} 1-(1-(e/c)^2)^3 & \text{if } |e| \leq c \\ 1 & \text{if } |e| \geq c \end{array} \right\}, \tag{M.4}$$

with derivative $\rho'(e\,;c) = 6\psi(\,e;c\,)/c^2$ where,

$$\psi(\,e;c\,) = e\left(1 - \big(e/c\big)^2\right)^2 \times I_{\{|e|\leq c\}}. \tag{M.5}$$

The indicator function I equals 1 if the expression inside the brackets is true and 0 otherwise. The constant c in ρ_1 equals 1.548 for an MM-estimator with a BDP of 50% and the constant c in ρ_2 (ψ_2) equals 4.685

for an MM-regression estimator with 95% efficiency. Let $\hat{\boldsymbol{\beta}}$ be an MM-estimator which is obtained by solving

$$\frac{1}{n}\sum_{i=1}^{n}\psi_2\left(\frac{y_i-\boldsymbol{X}_i\hat{\boldsymbol{\beta}}}{\hat{\sigma}}\right)\boldsymbol{X}_i=\mathbf{0}, \tag{M.6}$$

where $\hat{\sigma}$ is a scale S-estimate (Salibián-Barrera, 2005; Yohai, 1987). The scale S-estimate minimizes the M-scale $\hat{\sigma}(\boldsymbol{\beta})$ which for any $\boldsymbol{\beta}\in\mathbb{R}^p$ can be computed by solving

$$\frac{1}{n}\sum_{i=1}^{n}\rho_1\left(\frac{y_i-\boldsymbol{X}_i\boldsymbol{\beta}}{\hat{\sigma}(\boldsymbol{\beta})}\right)=\mathrm{b}, \tag{M.7}$$

where $\mathrm{b}=0.50$ to obtain a BDP of 50%. The S-regression estimator is the solution $\hat{\boldsymbol{\beta}}_{\boldsymbol{s}}$ such that $\hat{\sigma}=\hat{\sigma}(\hat{\boldsymbol{\beta}}_{\boldsymbol{s}})$. These S-regression estimates are used as initial values for $\hat{\boldsymbol{\beta}}$ in an iterative procedure to solve equation M.6. The constant c in ρ_1 equals 1.548 for an S/MM-estimator with a BDP of 50% and the constant c in ρ_2 (ψ_2) equals 4.685 for an MM-regression estimator with 95% efficiency.

Then, the $\bar{\mathrm{F}}_{\mathrm{mm}}$ test-statistic for hypothesis test Type A is given by

$$\bar{\mathrm{F}}^{\mathrm{A}}_{\mathrm{mm}}=\Big(\sum_{i}^{n}\rho_2(\bar{e}_i/\hat{\sigma})-\sum_{i}^{n}\rho_2(\tilde{e}_i/\hat{\sigma})\Big)/\hat{\lambda}, \tag{M.8}$$

where $\bar{e}_i=y_i-\boldsymbol{X}_i\bar{\boldsymbol{\beta}}$, $\tilde{e}_i=y_i-\boldsymbol{X}_i\tilde{\boldsymbol{\beta}}$ and let $\hat{\lambda}=2^{-1}(n-p)^{-1}\{\Sigma\psi_2^2(\hat{e}_i/\hat{\sigma})\}\{n^{-1}\Sigma\psi_2'(\hat{e}_i/\hat{\sigma})\}^{-1}$ be a standardizing constant, where $\hat{e}_i=y_i-\boldsymbol{X}_i\hat{\boldsymbol{\beta}}$. For hypothesis test Type B the test-statistic is given by

$$\bar{\mathrm{F}}^{\mathrm{B}}_{\mathrm{mm}}=\Big(\sum_{i}^{n}\rho_2(\tilde{e}_i/\hat{\sigma})-\sum_{i}^{n}\rho_2(\hat{e}_i/\hat{\sigma})\Big)/\hat{\lambda}. \tag{M.9}$$

M.3 How to compute the p-value

To obtain a p-value for hypothesis test Type A and hypothesis test Type B, we need to compute the probability that the test-statistic ($\bar{\mathrm{F}}$ and $\bar{\mathrm{F}}_{\mathrm{mm}}$) is at least as large as the observed value of the test-statistic, given

that the null-hypothesis is true. Since the test-statistic involves inequality constraints, its null distribution takes the form of mixtures of F-distributions. Only for a minimal number of problems closed form expressions for these mixing weights (also known as chi-bar-square weights) are known (Gouriéroux, Holly, & Monfort, 1982; Kudô, 1963; Shapiro, 1988). An intuitive way to think about the weights is the one parameter case. Under the null-hypothesis the parameter estimate $\hat{\beta}_1$ has an equal probability of 0.5 to be positive or negative. Under the scenario of a one-sided parameter constraint, e.g., $\hat{\beta}_1 > 0$, the test-statistic is under the null-hypothesis $F_{1,\nu}$ ($\nu = n - p$) distributed in 50% of the cases, when $\beta_1 > 0$, and equal to zero in the other 50% of the cases, when $\hat{\beta}_1 < 0$. Hence the null distribution is a mixture of 0 and $F_{1,\nu}$, with equal probability 0.5 (e.g., $0.5 \times 0 + 0.5 \times F_{1,\nu} = 0.5 \times F_{1,\nu}$). Fortunately, the mixing weights mixed over their degrees of freedom can be approximated sufficient precise by using the multivariate normal probability distribution function with additional Monte Carlo steps (Grömping, 2010) or the weights can be computed entirely by Monte Carlo simulation (Silvapulle & Sen, 2005; Wolak, 1989). In addition, the p-value can also be computed directly using bootstrapping (Silvapulle & Sen, 2005, pp. 78–81). By default, `restriktor` uses the multivariate normal distribution function with additional Monte Carlo steps. For more information about how to use the other methods, see `?iht`.

N
restriktor output

N.1 Output example 2

```
Restriktor: restricted hypothesis tests ( 36 residual degrees of freedom ):

Multiple R-squared reduced from 0.631 to 0.560

Constraint matrix:
   GroupNo GroupPhysical GroupBehavioral GroupBoth    op rhs active
1:       0             1              -1         0    ==   0    yes
2:      -1             1               0         0    >=   0     no
3:       0             0              -1         1    >=   0     no


Overview of all available hypothesis tests:

Global test: H0: all parameters are restricted to be equal (==)
         vs. HA: at least one inequality restriction is strictly true (>)
       Test statistic: 21.5452,   p-value: <0.0001

Type A test: H0: all restrictions are equalities (==)
         vs. HA: at least one inequality restriction is strictly true (>)
       Test statistic: 21.5452,   p-value: <0.0001
```

```
Type B test: H0: all restrictions hold in the population
         vs. HA: at least one restriction is violated
       Test statistic: 6.4857,   p-value: 0.06164

Note: All tests are based on a mixture of F-distributions
      (Type C test is not applicable because of equality restrictions)
```

N.2 Output example 3

```
Restriktor: restricted hypothesis tests ( 14 residual degrees of freedom ):


Multiple R-squared remains 0.985

Constraint matrix:
   GroupActive GroupNo GroupPassive    op rhs active
1:     -0.6596       0       0.6596    >= 0.2     no
2:           0  0.6596      -0.6596    >= 0.2     no


Overview of all available hypothesis tests:

Global test: H0: all parameters are restricted to be equal (==)
         vs. HA: at least one inequality restriction is strictly true (>)
       Test statistic: 3.1880,   p-value: 0.08858

Type A test: H0: all restrictions are equalities (==)
         vs. HA: at least one inequality restriction is strictly true (>)
       Test statistic: 3.1880,   p-value: 0.08858

Type B test: H0: all restrictions hold in the population
         vs. HA: at least one restriction is violated
       Test statistic: 0.0000,   p-value: 1

Type C test: H0: at least one restriction is false or active (==)
         vs. HA: all restrictions are strictly true (>)
       Test statistic: 0.7323,   p-value: 0.238

Note: Type C test is based on a t-distribution (one-sided),
      all other tests are based on a mixture of F-distributions.
```

N.3 Output example 4

```
Restriktor: restricted hypothesis tests ( 35 residual degrees of freedom ):

Multiple R-squared reduced from 0.685 to 0.609

Constraint matrix:
   GroupNo GroupPhysical GroupBehavioral GroupBoth Age_Z    op rhs active
1:       0             1              -1         0     0    ==   0    yes
2:      -1             1               0         0     0    >=   0     no
3:       0             0              -1         1     0    >=   0     no


Overview of all available hypothesis tests:

Global test: H0: all parameters are restricted to be equal (==)
         vs. HA: at least one inequality restriction is strictly true (>)
       Test statistic: 44.5941,   p-value: <0.0001

Type A test: H0: all restrictions are equalities (==)
         vs. HA: at least one inequality restriction is strictly true (>)
       Test statistic: 19.9963,   p-value: 0.0001172

Type B test: H0: all restrictions hold in the population
         vs. HA: at least one restriction is violated
       Test statistic: 8.4966,   p-value: 0.02751

Note: All tests are based on a mixture of F-distributions
      (Type C test is not applicable because of equality restrictions)
```

N.4 Output example 5

```
Restriktor: restricted hypothesis tests ( 16 residual degrees of freedom ):


Multiple R-squared remains 0.860

Constraint matrix:
   (Intercept) APS_Z Hours_Z Anxiety_Z    op rhs active
1:           0     1      -1         0    >=   0     no
2:           0     0       1        -1    >=   0     no
```

```
Overview of all available hypothesis tests:

Global test: H0: all parameters are restricted to be equal (==)
         vs. HA: at least one inequality restriction is strictly true (>)
       Test statistic: 98.4338,   p-value: <0.0001

Type A test: H0: all restrictions are equalities (==)
         vs. HA: at least one inequality restriction is strictly true (>)
       Test statistic: 12.3847,   p-value: 0.002534

Type B test: H0: all restrictions hold in the population
         vs. HA: at least one restriction is violated
       Test statistic: 0.0000,   p-value: 1

Type C test: H0: at least one restriction is false or active (==)
         vs. HA: all restrictions are strictly true (>)
       Test statistic: 0.4862,   p-value: 0.3167

Note: Type C test is based on a t-distribution (one-sided),
      all other tests are based on a mixture of F-distributions.
```

N.5 Output example 6

```
Restriktor: restricted hypothesis tests ( 269 residual degrees of freedom ):


Multiple R-squared remains 0.218

Constraint matrix:
   (Intercept) gender guilt anger TBSA age gender:guilt gender:anger
1:           0      0     0     0    0   0         2.02         2.06
2:           0      0     0     0    0   0         1.98         1.94
   gender:TBSA    op rhs active
1:        7.35    >=   0     no
2:       11.65    >=   0     no


Overview of all available hypothesis tests:

Global test: H0: all parameters are restricted to be equal (==)
         vs. HA: at least one inequality restriction is strictly true (>)
       Test statistic: 83.5889,   p-value: <0.0001

Type A test: H0: all restrictions are equalities (==)
         vs. HA: at least one inequality restriction is strictly true (>)
```

```
       Test statistic: 5.3064,   p-value: 0.04542

Type B test: H0: all restrictions hold in the population
         vs. HA: at least one restriction is violated
       Test statistic: 0.0000,   p-value: 1

Type C test: H0: at least one restriction is false or active (==)
         vs. HA: all restrictions are strictly true (>)
       Test statistic: 1.6422,   p-value: 0.05086

Note: Type C test is based on a t-distribution (one-sided),
      all other tests are based on a mixture of F-distributions.
```

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), (pp. 199–213). Springer New York: NY. doi: doi:10.1007/978-1-4612-1694-0_15

Baayen, C., Klugkist, I., & Mechsner, F. (2012). A test of order-constrained hypotheses for circular data with applications to human movement science. *Journal of Motor Behavior*, *44*(5), 351–363. doi: doi:10.1080/00222895.2012.709549

Bakker, A., Van der Heijden, P., Van Son, M., & Van Loey, N. (2013). Course of traumatic stress reactions in couples after a burn event to their young child. *Health Psychology*, *10*(32), 1076–1083. doi: doi:10.1037/a0033983

Bartholomew, D. (1961a). Ordered tests in the analysis of variance. *Biometrika*, *48*(3/4), 325–332. doi: doi:10.2307/2332754

Bartholomew, D. (1961b). A test of homogeneity of means under restricted alternatives. *Journal of the royal statistical society. Series B (Methodological)*, *23*(2), 239–281.

Berger, J., & Mortera, J. (1999). Default bayes factors for nonnested hypothesis testing. *Journal of the American Statistical Association*, *94*(446), 542–554.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997–1003. doi: doi:10.1037/0003-066X.49.12.997

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*(4), 286–300.

Egberts, M. R., van de Schoot, R., Boekelaar, A., Hendrickx, H., Geenen, R., & N.E.E., V. (2016). Child and adolescent internalizing and externalizing problems 12 months postburn: the potential role of preburn functioning, parental posttraumatic stress, and informant bias. *Child & Adolescent Psychiatry*, *25*(7), 791–803. doi: doi:10.1007/s00787-015-0788-z

Gouriéroux, C., Holly, A., & Monfort, A. (1982). Likelihood ratio test, wald test, and kuhn-tucker test in linear models with inequality constraints on the regression parameters. *Econometrica*, *50*, 63–80. doi: doi:10.2307/1912529

Grömping, U. (2010). Inference with linear equality and inequality con-

straints using R: The package ic.infer. *Journal of statistical software*, *33*, 1–31. doi: doi:10.18637/jss.v033.i10

Gu, X., Mulder, J., Deković, M., & Hoijtink, H. (2014). Bayesian evaluation of inequality constrained hypotheses. *Psychological Methods*, *19*(4), 511-527. doi: doi:10.1037/met0000017

Hoijtink, H. (2012). *Informative Hypotheses: Theory and Practice for Behavioral and Social Scientists*. Boca Raton, FL: Taylor & Francis.

Horowitz, M., Wilner, N., & Alvarez, W. (1979). Impact of event scale: a measure of subjective stress. *Psychosomatic Medicine*, *3*(41), 209–218.

Huber, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability, volume 1: Statistics* (pp. 221–233). Berkeley, Calif.: University of California Press. Retrieved from `http://projecteuclid.org/euclid.bsmsp/1200512988`

Huber, P. (1981). *Robust statistics*. New York: Wiley.

Huber, P., & Ronchetti, E. (2009). *Robust statistics*. New York: Wiley.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality Constrained Analysis Of Variance: A Bayesian Approach. *Psychological Methods*, *10*, 477–493. doi: doi:10.1037/1082-989X.10.4.477

Klugkist, I., van Wesel, F., & Bullens, J. (2011). Do we know what we test and do we test what we want to know? *International Journal of Behavioral Development*, *35*(6), 550–560. doi: doi:10.1177/0165025411425873

Kofler, M., Rapport, M., Sarver, D., Raiker, J., S.A., O., Friedman, L., & Kolomeyer, E. (2013). Reaction time variability in adhd: A meta-analytic review of 319 studies. *Clinical Psychology Review*, *33*(6), 795–811. doi: doi:10.1016/j.cpr.2013.06.001

Kudô, A. (1963). A multivariate analogue of the one-sided test. *Biometrika*, *50*(3/4), 403–418. doi: doi:10.2307/2333909

Kuiper, R. (2011). *Model selection criteria: How to evaluate order restrictions* (Dissertation, Utrecht University). Retrieved from `https://dspace.library.uu.nl/handle/1874/224499`

Kuiper, R., & Hoijtink, H. (2010). Comparisons of means using exploratory and confirmatory approaches. *Psychological Methods*, *15*(1), 69–86. doi: doi:10.1037/a0018720

Kuiper, R., Hoijtink, H., & Silvapulle, M. (2012). Generalization of the order-restricted information criterion for multivariate normal linear models. *Journal of Statistical Planning and Inference*, *142*(8),

42454–2463. doi: doi:10.1016/j.jspi.2012.03.007

Kuiper, R., Klugkist, I., & Hoijtink, H. (2010). A Fortran 90 Program for Confirmatory Analysis of Variance. *Journal of Statistical Software*, *34*, 1–31.

Maronna, R., Martin, D., & Yohai, V. (2006). *Robust statistics: Theory and methods*. John Willey and Sons, New York.

Mulder, J., Hoijtink, H., & Klugkist, I. (2010). Equality and inequality constrained multivariate linear models: Objective model selection using constrained posterior priors. *Journal of Statistical Planning and Inference*, *140*(4), 887–906. doi: doi:10.1016/j.jspi.2009.09.022

Nickerson, R. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301. doi: doi:10.1037/1082-989X.5.2.241

Perlman, M. (1969). One-sided testing problems in multivariate analysis. *The Annals of Mathematical Statistics*, *40*(2), 549–567. doi: doi:10.1214/aoms/1177697723

R Development Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. (ISBN 3-900051-07-0)

Richardson, M., & Abraham, C. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin*, *138*(2), 353–387. doi: doi:10.1037/a0026838

Roberts, N., Roberts, P., Jones, N., & Bisson, J. (2015). Psychological interventions for post-traumatic stress disorder and comorbid substance use disorder: A systematic review and meta-analysis. *Clinical Psychology Review*, *38*, 25–38. doi: doi:10.1016/j.cpr.2015.02.007

Robertson, T., Wright, F. T., & Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. New York: Wiley.

Rutherford, A. (2001). *Introducing ANOVA and ANCOVA. a GLM approach*. John Willey and Sons, London.

Salibián-Barrera, M. (2005). Estimating the p-values of robust tests for the linear model. *Journal of statistical planning and inference*, *128*(1), 241–257. doi: doi:10.1016/j.jspi.2003.09.033

SAS Institute Inc. (2008). Sas/stat® 9.2 user's guide [Computer software manual]. Cary, NC: SAS Institute Inc.

Schrader, R., & Hettmansperger, T. (1980). Robust analysis of variance based upon a likelihood ratio criterion. *Biometrika*, *67*(1), 93–101. doi: doi:10.2307/2335321

Shapiro, A. (1988). Towards a unified theory of inequality constrained

testing in multivariate analysis. *International Statistical Review*, *56*, 49–62. doi: doi:10.2307/1403361

Silvapulle, M. (1992). Robust tests of inequality constraints and one-sided hypotheses in the linear model. *Biometrika*, *79*(3), 621–630. doi: doi:10.2307/2336793

Silvapulle, M., & Sen, P. (2005). *Constrained statistical inference: Order, inequality, and shape restrictions*. Hoboken, NJ: Wiley.

Van de Schoot, R., Hoijtink, H., Mulder, J., Van Aken, M. A. G., Orobio de Castro, B., Meeus, W., & Romeijn, J. (2011). Evaluating Expectations about Negative Emotional States of Aggressive Boys using Bayesian Model Selection. *Developmental Psychology*, *47*, 203–212. doi: doi:10.1037/a0020957

Van de Schoot, R., & Strohmeier, D. (2011). Testing informative hypotheses in SEM increases power: An illustration contrasting classical hypothesis testing with a parametric bootstrap approach. *International Journal of Behavioral Development*, *35*, 180–190. doi: doi:10.1177/0165025410397432

Vanbrabant, L., Van de Schoot, R., & Rosseel, Y. (2015). Constrained statistical inference: sample-size tables for anova and regression. *Frontiers in Psychology*, *5*, 1–8. doi: doi:10.3389/fpsyg.2014.01565

W. N. Venables and B. D. Ripley. (2002). *Modern Applied Statistics with S*. Springer.

Wagenmakers, E. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Westfall, P., Tobias, R., & Wolfinger, R. (2011). *Multiple comparisons and multiple tests using SAS®* (Second ed.). SAS Institute Inc: Cary, NC.

White, H. (1980). "a heteroskedasticity-consistent covariance matrix and a direct test for heteroskedasticity". *Econometrica*, *48*, 817–838.

Wilcox, R. (2016). *Introducing to robust estimation & hypothesis testing* (4th ed.). Academic Press.

Wolak, F. (1987). An exact test for multiple inequality and equality constraints in the linear regression model. *Journal of the American statistical association*, *82*(399), 782–793. doi: doi:10.1080/01621459.1987.10478499

Wolak, F. (1989). Testing inequality constraints in linear econometric models. *Journal of Econometrics*, *41*(2), 205–235. doi: doi:10.1016/0304-4076(89)90094-8

Yohai, V. (1987). High breakdown-point and high efficiency robust estimates for regression. *The annals of statistics*, *15*(2), 642–656. doi:

doi:10.1214/aos/1176350366
Zelazo, P., Zelazo, N., & Kolb, S. (1972). "walking" in the newborn. *Science*, *176*, 314–315.

7 English summary

In the first study, presented in Chapter 2, we investigated the relation between sample-size reduction and order constraints. We showed sample-size tables at a prespecified power of 80% for order-constrained means in an ANOVA (e.g., $\mu_1 \leq \mu_2 \leq \mu_3$) and for positively-constrained regression coefficients in a linear model (e.g., $\beta_1 \geq 0, \beta_2 \geq 0, \beta_3 \geq 0$). The ANOVA results show that, depending on the number of groups involved, a maximum sample-size reduction between 30% to 50% can be gained when the full ordering between the means is taken into account. The linear regression results are comparable to the ANOVA results, but this only applies to the maximum number of constraints. In all other cases, the results show that an ordering of the parameters leads to a higher power compared to imposing positively constraints on the parameters. In addition, we showed that constraint misspecification has only a minor impact on the power.

In the second study, presented in Chapter 3, we investigated the performance of unconstrained, order- and positively-constrained OLS-, M- and MM-estimators when the data are contaminated with 10% bad leverage-points. The mean squared error (MSE) indicates that MM-estimation

produces the most precise estimates. For all estimators, it holds that the MSE improves most if the regression coefficients are subject to order constraints compared to positively-constrained coefficients and unconstrained coefficients. In addition, we investigated the size and power of order- and positively-constrained, and unconstrained robust and non-robust tests (likelihood ratio, Wald/F, and score). The results showed that all robust and non-robust tests are size accurate but that the robust tests need larger samples to maintain the nominal level. However, only MM-tests are capable of maintaining high power, where the robust likelihood ratio and Wald-test perform best. Again, the power improves most if the coefficients are subject to order constraints.

In the third study, presented in Chapter 4, we introduced the evaluation of an order-constrained hypothesis against its complement using the GORIC (weights). The GORIC is an information criterion that can be used to evaluate competing hypotheses in univariate and multivariate normal linear models, where the regression parameters are subject to order constraints. An individual GORIC value is not interpretable. To improve the interpretation, GORIC-weights and related evidence ratios can be computed. This ratio reflects the relative evidence for one hypothesis versus another. By means of a simulation study we demonstrated that the relative evidence for an order-constrained hypothesis against its complement increases for larger sample-size and/or effect-size, while the relative evidence for an order-constrained hypothesis against the unconstrained hypothesis has an upper-boundary.

In the fourth study, presented in Chapter 5, we described a general procedure for testing order-constrained hypotheses in structural equation models (SEM) using the `R` package **lavaan**. We used the likelihood ratio statistic to test constrained hypotheses and the resulting p-value was computed by either parametric or Bollen-Stine bootstrapping. Since the obtained p-value can be biased, a double bootstrap approach is available.

In the fifth study, presented in Chapter 6, we provided a tutorial for the `R` package **restriktor**. Restriktor can be used to estimate and evaluate informative hypotheses for the linear model. For seven examples, we showed how informative hypotheses could be evaluated using hypothesis testing and model selection using information criteria.

8 General discussion

8.1 Limitations and Further research

The power tables presented in Chapter 2 are based on the global $\bar{F}$-test in order to make a fair comparison to the frequently used classical (unconstrained) global F-test. This means that in all simulations the null-hypothesis for hypothesis test Type A is equal to the intercept-only model. For example, in an ANOVA with $k = 4$ groups and one order constraint, the null-hypothesis would equal $H_{A0} : \mu_1 = \mu_2 = \mu_3 = \mu_4$ and the alternative order-constrained hypothesis might equal $H_{A1} : \mu_1 < \mu_2, \mu_3, \mu_4$. Yet, in practice we are probably more interested in testing H_{A1} against $H_{A0} : \mu_1 = \mu_2, \mu_3, \mu_4$. We applied this latter approach in Chapter 3, where we also included the $\bar{F}$-test. Both methods showed a comparable relative decrease in sample-size but for the second approach larger samples are needed to maintain equal power.

In Chapter 2 and Chapter 3, we used hypothesis tests to find evidence in favor of an order-constrained hypothesis. One of the critiques of hypothesis testing in general is that it does not test the alternative hypothesis, hence it cannot be rejected or falsified. This means that we can

only find indirect evidence in favor of or against the hypothesis of interest. An analogue reasoning applies to hypothesis test Type A discussed in this dissertation. Rejection of hypothesis test Type A does not provide evidence for the imposed order constraints. The test concentrates its power as much as possible in the area where the constraints hold. Moreover, even with all constraints strictly (and statistically significant) violated its null-hypothesis can be rejected. Therefore, hypothesis test Type B plays a crucial role for providing evidence for the imposed order constraints. If we fail to reject the null-hypothesis of hypothesis test Type B, we would argue that we have found strong evidence that the imposed constraints hold in the data. Nevertheless, failing to reject hypothesis test Type B still does not mean that the order-constrained null-hypothesis is always true. If any of the constraints is violated, however small it may be, increasing the sample-size will eventually lead to rejection of hypothesis test Type B. Thus, although informative hypothesis testing provides us with a tool to test the order-constrained hypothesis directly, it cannot provide us with a wholehearted answer. Additional diagnostics (e.g., effect-size and a visual inspection of the parameters) are still required to strengthen the conclusion.

Clearly, the Neyman-Pearson approach leaves the possibility open that both the null and alternative hypotheses are invalid. Moreover, when there are two or more alternative hypotheses, hypothesis testing lacks ways to reject or falsify any of these alternative hypotheses. In this context, we should move forward from the historical methods to alternative methods. We argue that inferential data analysis should be based on the likelihood and related evidence ratios as discussed in Chapter 4. These methods do not need a formal null-hypothesis, test-statistic, significance level and p-value. Moreover, applying information criteria are so easy to both compute and understand, that researchers may be compelled to used them. However, model selection using information criteria - for order restrictions (GORIC) - is not ready yet for empirical data. This is because of several practical reasons. First, missing data are rule rather than the exception in social and behavioral research (Enders, 2003, 2010) but in contrast to hypothesis testing, literature is lacking on how to deal with missing values in case of model selection. Kuiper and Hoijtink (2011) have shown how information criteria such as the AIC can be computed in case of missing data but it is unclear if these results also account for order-

restricted information criteria. Second, the GORIC has been derived for univariate and multivariate normal linear models although empirical data always differ more or less from the assumed normality. Consequently it is questionable whether the GORIC is robust against such violations. Third, in many research fields problems with regard to sample-size arise. For example in studies with limited resources (e.g., in expensive fMRI studies), ethical issues (e.g., in case of vulnerable groups) or small populations (e.g., in clinical trials). The problem is that many standard statistical methods do not perform well anymore in case of small-samples. For the AIC, small-sample corrected versions have been developed (Hurvich & Tsai, 1989; Sugiura, 1978) but these do not yet exist for the GORIC.

In this dissertation, we developed the user-friendly (at least we tried) `R` package **restriktor** for estimating and evaluating order-constrained hypotheses. Although, we profoundly believe that restriktor is a welcome addition to an applied researcher's toolbox, it has some practical limitations. Missing data is not supported (yet). Restriktor is limited to linear models of class lm, rlm (robust), glm (generalized) and mlm (multivariate) and restriktor cannot handle nonlinear equality and/or inequality constraints. In addition, as mentioned above, small sample corrections as well as a robust version for the GORIC are missing. It is intended to tackle these limitations in the next few years. We recommend the interested reader to monitor the restriktor website at `www.restriktor.org` in order to be up to date.

8.2 Remaining issues

8.2.1 order-constrained variances

In this dissertation, we focused merely on one-sided means and regression coefficients. But instead of focusing on measures of central tendency, we can also focus on measures of dispersion, such as variances. Although, the problem of testing variances is well-known in the statistical literature, see e.g.,Molenberghs and Verbeke (2005, 2007, 2011) and Verbeke and Molenberghs (2000, 2003), much confusion remains on this matter. For example, in multilevel modeling (MLM) the intercept and slope coefficients are assumed to vary across clusters. In general, we will regard these cluster effects as random-effects. The variance components of these

random-effects are often of theoretical importance and hence often require inference on them. Whenever inference for variance components is required, the choice between one-sided and two-sided tests is inevitable and this choice depends on whether negative variance components are allowed. When negative variances are permitted, standard two-sided inferential procedures, such as the likelihood ratio and Wald test statistics can be used. When negative variances are not allowed, one-sided inferential procedures are necessary.

Probably as a result of the complex statistical literature, applied researchers often have misconceptions on the issue of testing variance components. At the heart of this confusion lies the fact that software defaults are often not well understood. The parameter space over which optimization is done, is key to whether a one-sided or two-sided tests is appropriate. It is often not known that this is determined by the software and often made implicitly by the software defaults used by a particular software package. For example, most typical MLM software procedures (e.g., HLM, lme, MLwiN) use constrained estimation by default, while most SEM software procedures (e.g., Mplus, lavaan, LISREL) use unconstrained estimation by default. This might suggest that admitting for negative variances lead to unaffected test statistics and that standard unconstrained inferential procedures can be employed. However, this is not the case. Consider for example the general linear mixed model (Laird & Ware, 1982) $\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$. If $\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$ are normally distributed (with mean zero, and variance $\boldsymbol{D}$ and $\boldsymbol{\Sigma}_i$), then the marginal distribution of $\boldsymbol{y}_i$ equals $N(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{V}_i)$, with $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \boldsymbol{\Sigma}_i$. If we only care about the marginal model, negative values for diagonal elements of $\boldsymbol{D}$ and $\boldsymbol{\Sigma}_i$ are perfectly acceptable, as long as $\boldsymbol{V}_i$ is positive definite. However, the positive definiteness of $\boldsymbol{V}_i$ imposes an implicit bound on the values of the diagonal elements. If they are too small, $\boldsymbol{V}_i$ may become negative definite, and therefore, we can not consider the variance parameters as completely unbounded. Consequently, the distribution of the variances are not symmetrical anymore and post-hoc adjustments are needed to obtain correct confidence intervals. Hence, we believe that testing variance components correctly remains an issue and that a clear tutorial paper is desired in which we list all points to test variance components correctly.

8.2.2 Partially adaptive estimation

In Chapter 3, we discussed robust estimation of the regression parameters as an alternative to OLS estimation if the data are contaminated with outliers. Robust estimators are obtained as the solution to $\min_b \sum_{i=1}^{n} \rho(y_i - X_i b, \hat{\eta})$, where $\rho(e)$ is a loss function less increasing than the squared loss in OLS and $\hat{\eta}$ is an estimate of the scale. While outliers may be the result of measurement errors, recording errors or other sources of errors, many outliers are actually generated by genuinely thick-tailed or asymmetric error distributions. During my PhD we encountered Partial adaptive estimation (Mcdonald & White, 1993) which allows for selecting an error distribution which includes unknown parameters (η) (e.g., scale, skewness and kurtosis) that can control the shape of the probability density function of the errors. To name a few distributions: the generalized error distribution (GED), the symmetric generalized t (GT), the skewed generalized t (SGT), the exponential generalized beta of the second kind (EGB2) and the inverse hyperbolic sine (IHS) (Hansen, McDonald, & Turley, 2006). These distributions form the basis for partially adaptive estimation.

While partially adaptive estimators (PAE) provide a powerful alternative to OLS in the presence of non-normal errors, they are completely unknown in the social and behavioral sciences. Hence, an accessible paper about constrained PAE and their applications is needed. In addition, it is noteworthy that the non-linear optimization problem associated with PAE is computationally more demanding then the linear optimization problem associated with OLS. The computational time increases for larger samples and the number of parameters. Moreover, we stumbled on convergence issues when trying to implement several PAE. Notwithstanding these current issues, partially adaptive estimation provide a flexible method to reduce unrealistic assumptions such as normality which underlie most methods for univariate and multivariate models.

8.3 Conclusion

In this dissertation, we focused on two alternative approaches to evaluate the hypothesis of interest more directly, i.e. informative hypothesis testing and model selection using order-restricted information criteria. These approaches have shown to be more 'powerful' than NHST. The main im-

plication is the possibility to reduce costs. Data collection in the social and behavioral sciences is usually the most expensive part of conducting research. Since the outcome of this dissertation ensures that researchers can use smaller samples, the costs of data collection can be reduced. In addition, researchers who are dealing with inevitable small samples in particular may benefit from these alternative approaches. Finally, we hope that this dissertation gives applied researchers a push to employ more informative hypotheses.

References

Enders, C. (2003). Using the expectation maximization algorithm to estimate coefficient alpha for scales with item-level missing data. *Psychological Methods*, *8*(3), 322–337. doi: doi:10.1037/1082-989X.8.3.322

Enders, C. (2010). *Applied missing data analysis*. The Guilford Press: New York, NY.

Hansen, J., McDonald, J., & Turley, R. (2006). Partially adaptive robust estimation of regression models and applications. *European Journal of Operational Research*, *170*(1), 132–143. doi: doi:10.1016/j.ejor.2004.06.008

Hurvich, C., & Tsai, C. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297–307. doi: doi:10.2307/2336663

Kuiper, R., & Hoijtink, H. (2011). How to handle missing data in regression models using information criteria. *Statistica Neerlandica*, *65*(4), 489–506. doi: doi:10.1111/j.1467-9574.2011.00496.x

Laird, N., & Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*(4), 963–974.

Mcdonald, J., & White, S. (1993). A comparison of some robust, adaptive, and partially adaptive estimators of regression models. *Econometric Reviews*, *12*, 103–124.

Molenberghs, G., & Verbeke, G. (2005). *Models for discrete longitudinal data*. Springer: New York.

Molenberghs, G., & Verbeke, G. (2007). Likelihood ratio, score, and wald tests in a constrained parameter space. *The American Statistician*, *61*, 22–27. doi: doi:10.1198/000313007X171322

Molenberghs, G., & Verbeke, G. (2011). A note on a hierarchical interpretation for negative variance components. *Statistical Modelling*, *11*(5), 389–408.

Sugiura, N. (1978). Further analysts of the data by akaike' s information criterion and the finite corrections. *Communications in Statistics - Theory and Methods*, *7*(1), 13–26. doi: doi:10.1080/03610927808827599

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Spring-Verlag: New York.

Verbeke, G., & Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics*, *59*, 254–262. doi: doi:10.1111/1541-0420.00032

9

Nederlandstalige samenvatting

In de eerste studie, gepresenteerd in hoofdstuk 2, hebben we de relatie tussen reductie in steekproefgrootte en orde-restricties onderzocht. We tonen tabellen met steekproefgroottes met een vooraf gespecificeerde power van 80% voor orde-gerestricteerde gemiddelden in een ANOVA (e.g., $\mu_1 \leq \mu_2 \leq \mu_3$) en voor positief-gerestricteerde regressiecoefficiënten in een lineair model (e.g., $\beta_1 \geq 0, \beta_2 \geq 0, \beta_3 \geq 0$). De ANOVA-resultaten tonen, afhankelijk van het aantal groepen, dat een maximale steekproefgrootte-reductie van 30% tot 50% behaald kan worden als de gemiddelden volledig geordend zijn. De resultaten voor de positief-gerestricteerde regressiecoefficiënten zijn vergelijkbaar met die van de ANOVA, maar dit geldt enkel voor het maximaal aantal restricties. In alle andere gevallen leidt een ordening van de parameters tot een hogere power dan het opleggen van positieve restricties. Verder blijkt dat voor beide testen kleine misspecificaties nauwelijks invloed hebben op de power.

In de tweede studie, gepresenteerd in hoofdstuk 3, hebben we de prestaties van orde-, positief- en niet-gerestricteerde OLS-, M- en MM-schatters onderzocht waarbij de data geïnfecteerd zijn met extreme waarden in zowel de uitkomstvariabele als in de onafhankelijke variabelen.

De resultaten tonen dat op basis van de gemiddelde-kwadratensom MM-schatters het meest accuraat zijn. Voor alle schatters geldt wel dat de gemiddelde-kwadratensom het kleinst is bij orde-gerestricteerde regressiecoefficiënten. Verder hebben we ook het nominale niveau en de power onderzocht van robuuste en niet-robuuste testen (likelihood ratio, Wald/F, en score). Uit de resultaten blijkt dat het nominale niveau van alle robuuste en niet-robuuste testen accuraat is, maar dat robuuste testen een grotere steekproefgrootte nodig hebben om het nominale niveau vast te houden. Voor de power geldt dat enkel MM-testen in staat zijn om een hoge power vast te houden. De robuuste likelihood ratio-test en de Wald-test presteren hierbij het best. Ook hier geldt dat de powerwinst het grootst is bij orde-restricties.

In de derde studie, gepresenteerd in hoofdstuk 4, introduceerden we een methode om orde-gerestricteerde hypothesen te evalueren tegen haar complement. Hiervoor maakten we gebruik van de GORIC. De GORIC is een informatiecriterium dat gebruikt kan worden om concurrerende informatieve hypothesen te evalueren in enkelvoudige of meervoudige lineaire regressiemodellen. Een GORIC-waarde op zichzelf is niet interpreteerbaar, maar het verschil tussen twee GORIC-waarden is wel belangrijk. Dit verschil kan namelijk vertaald worden naar een maat van relatieve evidentie. Deze ratio reflecteert de evidentie van de ene hypothese ten opzichte van de andere hypothese.

In de vierde studie, gepresenteerd in hoofdstuk 5, hebben we een algemene procedure voor het testen van informatieve hypothesen in structurele vergelijkingsmodellen geintroduceerd. Om informatieve hypothesen te toetsen hebben we gebruik gemaakt van de likelihood-ratiotest en de resulterende p-waarde kan door zowel de parametrische bootstrap als de Bollen-Stine bootstrap berekend worden. Vanwege het feit dat de p-waarde vertekend kan zijn, is er een dubbele-bootstrap-methode beschikbaar.

In de vijfde studie, gepresenteerd in hoofdstuk 6, beschreven we het `R`-pakket **restriktor**. Aan de hand van zeven voorbeelden lieten we zien hoe informatieve hypothesen geëvalueerd kunnen worden door hypothesetoetsing en door modelselectie op basis van informatiecriteria.

9.1 Conclusie

In deze dissertatie, lag de nadruk op twee alternatieve methoden om een hypothese te evalueren, i.e. informatieve hypothesetesten en modelselectie waarbij gebruik wordt gemaakt van orde-gerestricteerde informatiecriteria. Deze alternatieve methoden blijken 'krachtiger' (*more power*) dan NHST. Deze powerwinst impliceert dat de onderzoekskosten verlaagd kunnen worden. Dataverzameling in de sociale- en gedragswetenschappen is meestal de grootste kostenpost. Met de uitkomsten van deze dissertatie hebben we laten zien dat onderzoekers kleinere steekproeven kunnen gebruiken en dat daarmee de kosten van dataverzameling drastisch verlaagd kunnen worden. Bovendien zullen onderzoekers die te maken hebben met onvermijdelijk kleine steekproeven vooral profiteren van deze alternatieve methoden. Ten slotte, hopen we met deze dissertatie toegepaste onderzoekers een duwtje in de rug te geven om meer informatieve hypothesen te evalueren.

10
Data storage fact sheets

Data Storage Fact Sheet Chapter 2

% Name/identifier study % Author: Leonard Vanbrabant % Date: 29-05-2015

1. Contact details

1a. Main researcher

- name: Leonard Vanbrabant
- address: Henri Dunantlaan 1, B-9000 Ghent
- e-mail: leonard.vanbrabant@ugent.be

1b. Responsible Staff Member (ZAP)

- name: Yves Rosseel
- address: Henri Dunantlaan 1, B-9000 Ghent
- e-mail: yves.rosseel@ugent.be

If a response is not received when using the above contact details, please send an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies

- Reference of the publication in which the datasets are reported: Vanbrabant L, Van De Schoot R and Rosseel Y (2015) Constrained statistical inference: sample-size tables for ANOVA and regression. Front. Psychol. 5:1565. doi: 10.3389/fpsyg.2014.01565
- Which datasets in that publication does this sheet apply to?: all data

Data Storage Fact Sheet Chapter 3

% Name/identifier study % Author: Leonard Vanbrabant % Date: 22-11-2016

1. Contact details

1a. Main researcher

- name: Leonard Vanbrabant
- address: Henri Dunantlaan 1, B-9000 Ghent
- e-mail: leonard.vanbrabant@ugent.be

1b. Responsible Staff Member

- name: Yves Rosseel
- address: Henri Dunantlaan 1, B-9000 Ghent
- e-mail: yves.rosseel@ugent.be

If a response is not received when using the above contact details, please send an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies

- Reference of the publication in which the datasets are reported: Vanbrabant, L., Van De Schoot, R., Van Loey, N. and Rosseel Y. (2017) A General Procedure for Testing Inequality Constrained Hypotheses in SEM. Methodology.
- Which datasets in that publication does this sheet apply to?: all data

Data Storage Fact Sheet Chapter 4

% Name/identifier study % Author: Leonard Vanbrabant % Date: 22-11-2016

1. Contact details

1a. Main researcher

- name: Leonard Vanbrabant
- address: Henri Dunantlaan 1, B-9000 Ghent
- e-mail: leonard.vanbrabant@ugent.be

1b. Responsible Staff Member

- name: Yves Rosseel
- address: Henri Dunantlaan 1, B-9000 Ghent
- e-mail: yves.rosseel@ugent.be

If a response is not received when using the above contact details, please send an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies

- Reference of the publication in which the datasets are reported: Vanbrabant, L., Van Aelst, S., Egberts, M., van de Schoot, R., and Rosseel, Y. Comparing inequality-constrained robust and non-robust regression estimation methods for one-sided hypotheses.
- Which datasets in that publication does this sheet apply to?: all data

Data Storage Fact Sheet Chapter 5

% Name/identifier study % Author: Leonard Vanbrabant % Date: 22-11-2016

1. Contact details

1a. Main researcher

- name: Leonard Vanbrabant
- address: Henri Dunantlaan 1, B-9000 Ghent
- e-mail: leonard.vanbrabant@ugent.be

1b. Responsible Staff Member

- name: Yves Rosseel
- address: Henri Dunantlaan 1, B-9000 Ghent
- e-mail: yves.rosseel@ugent.be

If a response is not received when using the above contact details, please send an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies

- Reference of the publication in which the datasets are reported: Vanbrabant, L., Van Loey, N., & Kuiper, R. Giving the complement a compliment: Evaluating an order-constrained hypothesis against its complement using the GORIC.
- Which datasets in that publication does this sheet apply to?: all data

Data Storage Fact Sheet Chapter 6

% Name/identifier study % Author: Leonard Vanbrabant % Date: 22-11-2016

1. Contact details

1a. Main researcher

- name: Leonard Vanbrabant
- address: Henri Dunantlaan 1, B-9000 Ghent
- e-mail: leonard.vanbrabant@ugent.be

1b. Responsible Staff Member

- name: Yves Rosseel
- address: Henri Dunantlaan 1, B-9000 Ghent
- e-mail: yves.rosseel@ugent.be

If a response is not received when using the above contact details, please send an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies

- Reference of the publication in which the datasets are reported: Vanbrabant, L., Van De Schoot, R., and Rosseel Y. An introduction to restriktor: informative hypothesis testing for AN(C)OVA and linear models.
- Which datasets in that publication does this sheet apply to?: all data