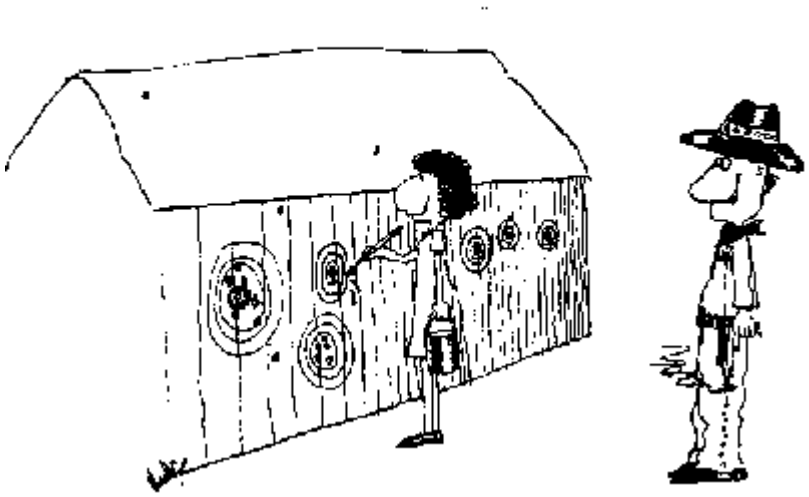
Data Mining:
Vice or Virtue?



Willem van der Deijl

Erasmus Institute for Philosophy and Economics,
Erasmus University Rotterdam

Research Master Thesis by Willem van der Deijl

Student Number: 36223410

Supervisor: Prof. Dr. J.M. Reiss

Advisor: Dr. H.C.K. Heilmann

Third reader: Prof. Dr. J.J. Vromen

Date of Completion: 05-08-2013

Word Count: 38,000

Preface and Acknowledgments

Before you lies the end result of studying two years at the Erasmus Institute for Philosophy and Economics (EIPE) and a year of thinking about this topic and couple of months of hard work. The topic of this thesis is statistical inference in economics. To some this may sound like a dull topic for something to work on for a couple of months. However, I have enjoyed it greatly. Ideas for this topic I developed in the courses of Julian Reiss in my first year at EIPE, but have been with me for a while. Statistical inference is something that has fascinated me for a number of years now. Besides a natural fascination I have for this topic, there are also a number of other reasons for caring about the topic. One of which is that statistical inference is one of the main ways in which economists study economic reality. This is something that has been stressed by Deirde McCloskey and Stephen Ziliak in work published over the last couple of decades. Their work has inspired me. However, the topic of this thesis is very different from theirs. One conclusion though, is similar. The economics profession may pay too much attention in their work to a single concept: p-values. Economics is a very important field of study whose insights may affect the welfare of billions of people. It is therefore important how we can learn from economic data.

Another reason is that it is a very small contribution to a very large philosophical discussion. Many philosophers have pondered over the question of how we can make sense of the world. While these problems at large are not addressed specifically in this thesis, the topic of this thesis is one of the many challenges for an empirical approach to understanding the world. I hope this thesis, however small, can be considered a contribution to this project.

I am grateful to many people that have supported me in many different ways over the last two years in general and the last months in particular. In particular I owe many thanks to Julian Reiss, my supervisor, and Conrad Heilmann, my advisor. At the early stages of developing my thesis, Julian Reiss, who helped me develop the ideas presented in this thesis, moved to Durham University. The fact that I could still go through with this thesis is due to the fact that Julian Reiss was very flexible, met me whenever he was in Rotterdam and always responded very quickly to drafts. At the same time, Conrad Heilmann has done much more for the development of this thesis than the position of advisor required him too. He helped me to restrict the topic of this thesis into the appropriate scope and took hours of his time to talk to me about thesis writing and other important academic-related choices I have had to make over the past months. I want to thank them both for this.

I also have to thank many friends and family members for understanding that I have had an occupied mind, and a busy schedule, for the past couple of months. Special thanks are due to Nina Kloeg, who has been with me in this process, and helped me rewrite the final draft. I also want to thank my parents, Christina and Leenderd van der Deijl, both for emotional and financial support over the last five years. Without their help, I would not have been able to write this thesis.

My experience at EIPE over the past two years has been great, and for that I have to thank everybody at EIPE, but in particular my fellow students Darian Heim, Philippe

Verreault-Julien, and Vaios Koliofotis. As the most junior and least experienced member of the group I have been able to learn a lot from them over the past two years. Furthermore I want to thank Darian Heim, Fred Muller, and Melissa Vergara Fernández in particular, and EIPE in general for very helpful comments at an early paper on this topic presented at the EIPE seminar in December, 2012. Furthermore, a version of chapter 3 was presented at the Philosophy of Science in a Forest conference in Leusden, May, 2013, and a version of chapter 4 was presented at the INEM conference in Rotterdam, June, 2013. In both cases, the audiences provided very helpful comments. The idea for the interlude to chapter 4, about the Look Elsewhere Effect in the Higgs boson discovery comes from my brother Pieter van der Deijl, who worked himself on the project, and told me about the way they dealt with multiple testing. This was a great inspiration for this thesis.

Contents

Chapter 1: Introduction

1.1. Introduction

In 1997, the American Economic Review published a seven-page paper with the title *I Just Ran Two Million Regressions* (Sala-i-Martin, 1997). A regression is a statistical model. Running two million regressions is estimating two million statistical models on the basis of the data. This large amount of estimations was conducted in order to finally find an answer to the question: What causes growth[1]? Soon, this became a widely discussed paper in the economics literature. "Surely, running that many regressions is not going to help us much", was a common reaction (e.g. Durlauf and Quah, 1999). Not everyone shared this sceptical attitude. Sala-i-Martin et al. (2004) continued the line of research by estimating 89 million regressions; Steel and Ley (1999) estimated over two trillion.

This is probably the most widely discussed case of data mining in the field of economics. One of the reasons it is widely discussed is not only that data mining was conducted on a large scale, but also that it was done so openly, almost proudly. The concept of data mining had always been a term for bad practice. Consider, for instance, some other cases that are called data mining. Most of them are considered bad or, at least, not completely 'kosher'.

Example 1: A researcher knows from theory that the best way to model the central bank lending rate is to model it as a function of the rates in former periods. She does not yet know though, how many periods in the past are actually relevant for the determination of the lending rate (see Liew, 2004, for a discussion on this problem). A variable in a past period used to explain contemporary values of the same variable is called a lag. In order to determine how many lags she should include in the model, she runs a number of tests and uses the one that fits best.

Example 2 is borrowed from Keynes' (1939) discussion of Tinbergen's (1939) paper on business cycles. A researcher has two data sets available. She examines both of them, but only uses the one that fits the model best. Keynes accuses Tinbergen of doing this.

Example 3: A researcher suspects that there may be unexplained patterns in the data. In order to detect such patterns, she runs a number of statistical tests that provide an indication of whether this is the case or not. She finds evidence of unexplained patterns in the data and respecifies the model in order to find a more comprehensive model.

Cases of data mining fall in three categories. Firstly, there are cases of "trying out" different models, or hypotheses, to see which one fits the data best. This can be done on a large scale (e.g. Sala-i-Martin, 1997). Or, it can be done on a small scale as the first example shows. In applied research it often occurs that the specific form of the model under research is not known completely. In such cases comparing different models based on their statistical properties seems an obvious solution. Doing so would be data mining of the first type. Secondly, data mining can be the selection of data to a pre-formed hypothesis. This however, is more likely to happen in experimental sciences than in economics. One example of this in

[1] The method for doing this was not as simple as it might seem from this description. See chapter 3.

economics is selection of sub-sample periods in such a way that the data fits the hypothesis better. Keynes' accusation of Tinbergen, example 2, is one such example. Lastly, there is the case of diagnostics testing. There are a number of features of econometric models besides model fit that may indicate whether a model is correctly specified. Testing whether these features are present, by means of statistical methods, is called diagnostic testing. This happens in example 3.

As the examples show, there is quite a large array of instances that are called data mining. Generally we can say that data mining lies at the extreme inductive end of the spectrum of inferences. Very roughly speaking, data mining occurs when the answer to a question is not sought by use of much theory, but by looking at what the data tell us. A related term is post hoc, or ad hoc, inference, which means that a theory explaining the data is constructed only after observing a number of relevant features of the data. Data mining is always an instance of post hoc inference, but has a more specific meaning: constructing a theory after having examined the statistical relationships present in the data. The meaning of the term will be described in more detail below.

The attitude towards data mining is generally very negative (cf. Leamer, 1978; Hoover and Perez, 2000), though some instances, such as diagnostic testing, are considered more acceptable than others (Spanos, 2000). Data mining ensures that a model is found that looks very good in terms of statistical criteria, even if it is completely wrong. The correct way to do econometrics, so it is thought, is to formulate a model first, and then estimate it on the data. In this case, statistical testing of the relationships is appropriate. If one formulates a model independent of a certain data set, which fits almost perfectly on this data set, this appears to be a very good indication that the model is correct. If it would not be, it is very exceptional that it fits so good nevertheless. However, if the data set is already used to formulate a model with the purpose of fitting this data set, the fact that the model indeed fits with the data is not surprising at all. It would fit regardless of how accurate the model is. The reader of an economics paper cannot detect, without more information, whether a result was brought about by data mining or not. Data mining is thus not just a problem in and of itself. It makes people doubt whether econometric results in general are scientific, or unreliable alchemy (see Hendry, 1980). Data mining does not only have a bad name, it gives econometrics as a whole a bad name, too.

At the same time, it has been observed that data mining is common practice (Pagan and Veall, 2000; Mayer, 2000). In research conducted by Burger and Du Plessis (2006), the researchers examined 75 papers that were published in renowned economics journals and found that 89% of them contained research practices that fell under their definition of data mining. In the methodological discussions on this topic, biblical references to sins became common to describe the general attitude in economics towards the practice (starting with Leamer, 1978, but also in Keuzenkamp, 1995; Hoover and Perez, 2000; Kennedy, 2002; Hendry, 2002; Magnus, 2002). It appears it was strongly felt that data mining is bad practice, but very tempting too: it guarantees publishable results without having to lie about the data. One only has to omit part of the procedure that resulted in the final model. We can call the view that data mining is bad and should be avoided the *Condemnation Thesis* (henceforth **CT**). This view seems as widely accepted as the view that data mining is widespread. Textbooks (e.g. , Gujarati, 2003; Verbeek, 2006; Wooldrige, 2009) preach caution, while at

the same time the American Economic Review publishes Sala-i-Martin's paper (1997). Its title proudly announced giving in to sin. In short, many acknowledge that data mining is somewhat problematic, yet it occurs regularly.

The methodological ambiguity does not stop here. Two further theses have been advanced about data mining that are hard to reconcile with **CT**. Firstly, it has been defended that data mining is desirable (Greene, 2000; Hoover and Perez, 2000; 2004; Campos et al., 2005). Data mining is studying the data carefully, and learning from it. This is arguably a very important research practice. This is especially the case when little theory is available to guide empirical research. A second thesis is that there is something unavoidable about data mining, particularly in the context of economic, non-experimental, data (e.g. Leamer, 1978; Denton, 1985; Caudill, 1990; White, 2000; Greene, 2000; Hoover and Perez, 2000). That is, even if individual researchers do not purposefully mine the data themselves, there is bound to be an evolutionary process in economics with regards to which hypotheses survive all the statistical testing and which ones do not. The surviving hypotheses will be reused in empirical research and theory, but may capture random empirical regularities rather than genuine relationships. Hence, according to them, there is data selection of hypotheses, even if we do not deliberately engage in data mining.

In short, besides the claim that data mining is bad methodological practice, it has been argued that data mining is widely practiced, is in fact unavoidable, and is desirable. There is a strong tension between these claims and **CT**. Perhaps not surprisingly, there are alternative views. Most notably, Edward Leamer has defended a Bayesian approach to data mining that does not condemn it at all. Also, David Hendry has defended a methodology of econometrics, the LSE approach, which embraces data mining. Hoover and Perez (1999) developed a method to apply their approach, the general-to-specific approach (henceforth *Gets*). This approach was further developed by Hendry and Krolzig (1999), and defended by Hoover and Perez (2000; 2004) and Hendry and Krolzig (2004). In addition to the fact that data mining may be desired, unavoidable and widely practiced, these authors argue that data mining, if done correctly, is in fact methodologically sound (Leamer, 1978; Hoover and Perez, 1999; 2000; 2004; Hendry, 2002; Hendry and Krolzig, 2004). In fact, Hoover and Perez (1999) and Hendry and Krolzig (2003) argue that their methodology may in fact be better than conventional methods.

The goal of this thesis is to sort out these different claims about data mining in order to come to a conclusion about the methodological consequences of the topic. The main question I tackle is: What are the methodological consequences of data mining? Unfortunately, my answer, in the end, shall not be that it is good, or that it is bad. I argue that the two main problems associated with data mining do not offer good reasons for condemning it, or for distinguishing good from bad practices. These two problems will be briefly discussed in this chapter and critically assessed in chapter 3 and 4. The first one is that data mining is (almost always) double counting: using evidence in both the construction and assessment of hypotheses, or theories (see chapter 3). The second is that the multiple testing involved in data mining distorts the inferential interpretation of statistics (see chapter 4 and 5). The conclusion that these problems do not offer good dimensions to assess the degree to which data mining practices are problematic may be seen as a somewhat negative conclusion. However, the overall message of the thesis is also that the skeptical **CT** is very implausible. Hence, my

thesis opens up the space for warranted data mining in some sense. Unfortunately, I cannot conclude that data mining is good methodological practice. There are simply many problems involved that do not yet have clear solutions. To some extent, the degree to which they cause a problem remains a black box. However, I do hope to have made clear by the end of this thesis what the harms are, and why data mining remains a large methodological question mark.

In this chapter I focus on a simpler question: What is data mining? In order to provide an answer, I present a historical overview of the methodological discussion on data mining, a discussion on the definition and describe two major problems for data mining.

1.2. Data mining: a brief history of the debate in economics

Early econometrics and data mining

While the term data mining was not used yet, the first discussion of data mining as a methodological problem is John Maynard Keynes' (1939; 1940) well-known critique of Jan Tinbergen's Statistical Testing of Business-cycle Theories I (1939; cf. Campos et al., 2005). Tinbergen had constructed one of the first econometric models to describe the economy of the United States. In order to decide what kinds of variables should be included in the model and how important they were, he studied the data. When a variable correlated highly with another, he included it in the model. He ran many tests to see if the model was good. Keynes was very sceptical about this methodology. He listed a number of concerns he had with the statistical methodology that could not easily be resolved. He holds an extremely sceptical view of the statistical method, ending on the note "Newton, Boyle and Locke all played with Alchemy. So let [Tinbergen] continue" (1940, p. 156). Two of the problems Keynes lists with respect to Tinbergen's methodology would fall under the header of data mining accusations in current terminology. Firstly, Tinbergen is accused of selecting the sub-sample for study in such a way that it would best fit the data (type-2 data mining in the typology discussed above). Secondly, he is accused of selecting the determinants of industrial production (his main variable of interest) by means of first examining their correlation with industrial production (data mining type 1). Keynes argued that the "trial-and-error" method that Tinbergen uses to select the number of lags he believes should be included in the statistical model, does not give much confidence in whether these lags are actually the relevant ones. This point was also taken up by Milton Friedman's discussion of Tinbergen's approach (Friedman, 1940). He argues that the statistical success of the models from his method is tautological, and has little to do with their adequacy:

"…the leads and lags, and various other aspects of the equations besides the particular values of the parameters (which alone can be tested by the usual statistical technique) have been selected after an extensive process of trial and error because they yield high coefficients of correlation. Tinbergen is seldom satisfied with a correlation coefficient less than 0.98. But these attractive correlation coefficients create no presumption that the relationships they describe will hold in the future."[2] (Friedman 1940, p. 659).

[2] Prediction, in this context is a reality check for Friedman.

The danger of data mining, or post hoc inference, is also taken up by another influential early econometrician: Trygve Haavelmo (1944). He discusses the problem that there are infinitely many mathematical forms that may fit any set of data (he discussed this in chapter 5 of his doctoral thesis). An applied econometrician may search for a function that fits the data, but he has to realize that the fact that a mathematical function is found that fits the data is by no means impressive. And therefore, it is important not to take fit with the data by itself as an argument for the truth of the hypothesis that is being tested. This worry comes very close to the Duhem problem as discussed by John Worrall (2006; 2010), and will be discussed in chapter 3 of this thesis.

Keynes' critique, which is much more sceptical than Friedman's and Haavelmo's, was received moderately well at best. Many thought that Keynes was hindering the progress of statistical analysis unnecessarily (cf. Louca, 2007), or simply did not understand the subject matter particularly well (cf. Hendry and Morgan, 1995). Moreover, it did not alter the enthusiasm for the newly developed methodology all too much. After all, econometrics took off as an important research method and ended up in the core of economics today.

Probably more influential was an alternative critique, written by Koopmans (1947). He did not raise any concerns with probabilistic interpretations of statistics in case of post hoc inference, but raised a worry about empirical inference without theory. Koopmans observed that Arthur Burns and Wesley Mitchell (1946) were effectively searching for explanations of business cycles in the data without the use of theory. He argued that this could not lead to solid causal inferences. Koopmans raised three problems with respect to empirical approaches that make no (or very limited) use of theory: 1) without theory, one does not know where to expect economic relationships: one is looking the dark, 2) without theory, causal relations cannot be inferred from the data: theory-poor research will therefore be of limited use to policy makers, and 3) in order detect relationships one has to know the structure behind them, which can only be done by means of a theoretical understanding of the phenomena. The second problem is not specific to data mining practices, and is not necessarily an argument against it, as data mining means that no theory is used in the discovery of the statistical relationship, but not necessarily that no theory was used in the interpretation thereof. The first and the third problem are related to the discovery of statistical relationships, and are therefore relevant problems for data mining in general. However, in practice, data mining is never a method of inference in which no theory is used whatsoever. In the cases discussed in this thesis, there is generally usage of background theory to sort out the right places to look for economically important statistical relationships. For instance, in example 1, discussed above, background theory was used to determine that one has to look at the lagged values of the central bank lending rate in order to explain the current values. However, the data was mined in order to find out which specific lagged values worked best. Koopmans' worries are very relevant warnings for methods in which there is much uncertainty about background theory, which is something we shall discuss much in chapter 2. In short, Koopmans raises some important points of concern for methods of inference that use little theory, but the specific issues it raised are not necessarily problems for data mining. What is likely most important about Koopmans' paper is its effect on the economic profession. Blind data research got a bad name.

While Keynes' critique did not have the impact it may have intended – to nip statistical economics in the bud – he may have consented to the fact that econometrics preceded on a hypothetico-deductive path. For a long time it was consensus in the economics profession that while data can test a theory, it is wrong to use data to form and test a theory, which is effectively what Tinbergen did. According to Roger Backhouse and Mary Morgan (2000), **CT** was established in these early stages, and was greatly influenced by the conventional view in philosophy of science at the time (also in Keuzenkamp, 1995). This view asserted a distinction between theories and observations to which they were tested. A deviation from this methodology deserved harsh treatment. Data mining was such a deviation.

Modern data mining debates

Important criticisms of the status quo view of data mining as bad practice were published in the late 1970's. Edward Leamer's book *Specification Searches* (1978) and David Hendry's inaugural lecture *Econometrics: Alchemy or Science* (1980) started a reconsideration of the standard econometrics approach. Both Leamer and Hendry developed radical alternative views.

David Hendry discussed Keynes' critique and wondered to what extent his criticisms of Tinbergen's work were still applicable to the econometrics at the time. His answer was that much work had been done in the econometrics profession to deal with the logical traps of statistical reasoning of which Keynes warned the economics profession. However, they remained relevant. He added another set of problems to Keynes' list with which econometrics has to deal. He wondered if econometrics was to be seen as a science or alchemy. On the one hand, he argued, it could be seen as alchemy: a researcher who is persistent can always find the right kind of evidence for his claims. On the other hand, proper methodology could turn econometrics into a science. He expressed hope that econometricians could deal with many of the issues by means of his famous dictum: "test, test and test" (p.403). While Campos et al. (2005; Hendry is a co-author) admits that in 1980 econometrics was not yet developed enough to deal with all the problems Keynes (1939) addressed, Hendry's (1980) overall attitude was that Keynes' accusation of econometrics being alchemy was grossly exaggerated simply because econometricians do much more than simply correlation fitting and hypothesis testing, at which the criticisms of Keynes and Friedman were aimed. Taking into account the large battery of tests that econometricians run on the correlations that they find, these are to be considered much more reliable than Keynes makes them appear.

Leamer's book challenged the condemnation of researchers who use data to construct econometric models. In fact, he argues that there is something unavoidable about this, and that it therefore makes little sense to condemn. Immanual Kant famously argued that ought implies can. If something cannot be done, we cannot maintain that people nevertheless need to do it. If data mining is unavoidable, we cannot expect researchers to avoid it. According to Leamer, the best way to go about the matter of data mining is to establish some rules to ensure that the final model does indeed yield useful results[3]. Key in his approach to data mining is *robustness*. Robustness means that a finding does not only hold in one particular

[3] For Leamer, it is important that there is no such thing as a true model, or a data generating process. Leamer is an instrumentalist about economic models, which had led to some disagreements with others working in this field.

setting, but holds even if the assumptions underlying it changes. The setting can refer to the data set in which it is tested, but also the model that is used to test a correlation. Leamer was particularly concerned with the second (Leamer, 1985). Maybe a simple statistical correlation is not so relevant if it was brought about by a search procedure, but if one variable is significantly correlated to another in a large variety of statistical models, it is very useful and relevant knowledge. If, on the other hand, it sometimes leads to insignificant coefficients (or coefficients that are inconsistent with coefficients in other models) it is a *fragile* finding. Revolutionary in Leamer is the fact that he endorses inference in which the data is used to form theories, and thereby goes against the hypothetico-deductive view of science that was one of the roots of data mining scepticism (Backhouse and Morgen, 2000). As an alternative to the received view he proposes a "Sherlock Holmes approach" to applied research. In *A Study in Pink*, Sherlock Holmes warns his friend Dr. Watson for a non-empirical approach to science: "It is a capital mistake to theorize before you have all the evidence. It biases judgment." Leamer uses this dictum to describe his views on empirical research in economics.

While Leamer and Hendry gave a great impulse to open up the debate with respect to the methodological rules considering statistical inference in economics, there was a renewed attention to the problems of data mining as well. Michael Lovell (1983) published an important paper to highlight the severity of the problems of data mining. In this paper an experiment on an artificial data set with random variables was conducted, and it was shown that the reliability of the statistical inferences greatly decreased the more data was mined. His paper provided a strong argument for caution with respect to data mining. This point was taken up by Frank Denton (1985; Caudill, 1988), who argued that data mining is not just a problem for individual researchers, but it may affect the results of a group of researchers working on the same data (as I discuss in chapter 4).

The renewed interest in data mining in the methodological debate was reflected in a renewed interest for the topic in the field of growth economics in the 1990's. A number of researchers started to apply versions of Leamer's methodology in their searches for a model to explain growth (beginning with Levine and Renalt, 1992). Sala-i-Martin's paper was on the same topic. The LSE approach methodologists took part in this debate as well (Hoover and Perez, 2004; Hendry and Krolzig, 2004). Hoover and Perez (2004) performed simulation experiments on growth economics data to argue that their approach performs better than all the other methods doing this. This started a new debate among philosophers and economists (See the special edition on the topic in the *Journal of Economic Methodology*, 7(2), 2000).

A key factor in the debates on data mining in the last 20 years is the fact that it becomes easier by the year to compute very complicated search procedures (e.g. Glymour et al., 1997). While Hoover and Perez (2004) report that running their automated search procedure takes a day and a half per search, Hendry conducted the same search procedure in a matter of seconds live in a talk he gave at the Institute for New Economic Thinking in 2010. The developments in computer technology have triggered much research on the matter and many new search algorithms have been developed over the past decade (DuPlessis, 2009). The methods designed for data analysis are strictly intended for dealing with statistical analysis of models that are estimated once (e.g. Wooldridge, 2009; and Hollanders, 2011). At the same time, the ease by which we can estimate equations is something that should be very beneficial for the subject and should be something we can learn from. As the classical

techniques of statistical inference do not allow us to do this, a new discussion on the methodology of data mining becomes increasingly important.

1.3. Definition and distinctions

In the preceding parts of this chapter I have described a number of instances of data mining and explained the meaning of the term in a general manner. For purposes of clarity, it is important to be more precise. As I focus on data mining in economics, which is mostly a non-experimental science, I do not consider type 2 data mining: fitting the evidence to a fixed hypotheses, and focus mostly on data mining of the first type: fitting hypotheses to the available evidence.

1.3.1. Definition

What many of the definitions found in the literature have in common is that data mining is considered something that goes against a standard view of the methodology of econometrics. For instance, Aris Spanos writes: "The term 'data mining' is usually used derisively to describe a group of activities in empirical modelling that are considered to lie outside the norms of 'proper' modelling." (Spanos, 2000). Or, as Du Plessis (2009) puts it: "(…) data mining seems to offend against norms of good conduct in econometrics; or at least, such norms were widely shared until recent advances in the theory of econometric modelling". (p.3). It is therefore useful to say something about what this received view on methodology is. To do this is to provide a somewhat stylized account. Spanos, for instance, stresses that there are very few economists who subscribe exactly to what is taught in textbooks, and argues that the received view is mostly a straw man. Moreover, even textbooks add critical discussions to their discussions of "proper" methodology (Gujarati, 2003; Verbeek, 2008; Wooldridge, 2009). However, as data mining is often contrasted with such a received view, it is useful to discuss the accounts that some give of this standard approach, even if it is a straw man. Jan Magnus (1999) and Hoover and Perez (2000) provide a useful discussion on what they take to be the received view on econometric methodology.

A crucial part of this received view of econometrics is that econometrics can test theories, but should not be used to develop them. There should be a strong separation between theory development and theory testing, and theory development should come first. Hoover and Perez (2000), for instance, put it as follows: "Only a well specified model should be estimated and if it fails to support the hypothesis, it fails; and the economist should not search for a better specification." Data mining is a deviation from this process (Hoover and Perez, 2000). The received view seen as such is clearly influenced by Koopmans' (1947) plea for theory based econometrics, and the hypothetico-deductive spirit that ruled among the early econometricians in its influential founding years (Backhouse and Morgan, 2000).

Some authors argue that data mining simply means a deviation from the received view. Data mining is what occurs when data is used in the formation of theories. Chris Chatfield, for instance, stresses this: "… models are not fully specified a priori, but rather are formulated, at least partially, by looking at the same data as those later used to fit the model"

(Chatfield, 1995, p. 426, as quoted by Du Plessis, 2009). This aspect of data mining is referred to by Du Plessis (2009) as dual usage of data, which is closely related to double counting (see chapter 3). Du Plessis takes this to be key in the definition of data mining. While I take data mining to always be an instance of using data in the formation of hypotheses and the evaluation thereof, it is not necessary for data mining that the evaluation and formation of hypotheses occurs with the same data. This is a fine distinction not made by Du Plessis. Using the same data for the formation and evaluation of hypotheses is both double counting and data mining. Using different data for the formulation of hypothesis and evaluation of hypothesis in order to maximize success of the model in terms of statistical criteria used, is data mining, but not double counting (an example of this methodology is RETINA, a method discussed in more detail in Chapter 3).

This description of data mining implies that any way in which a researcher is influenced by the data in the formulation of the hypothesis is a form of data mining. In its most common form one fits a number of models to the data to find out which one works best. This narrow description of data mining is also taken by some to define data mining. Gujarati, for instance, defines data mining as: "trying every possible model with the hope that at least one will fit the data well." (2003, p. 74). If formulated in this way, data mining is something that is hard to avoid in applied work. For instance, I discussed the case of the researcher who does not know how many lags of an independent variable are needed for a correct model, and in order to find out he estimates all models to see which one fits best.

Diagnostic testing in order to detect misspecifications that may help to correct the model is a way in which model formation is influenced by the data and is considered data mining by some (cf. Spanos, 2000). Some definitions of data mining therefore go as far as this. Thomas Mayer, for instance, writes "Within the broadest definition, data mining is the fitting of more than one econometric specification of the hypothesis" (2000, p. 184).

A widely used and more sophisticated account is that of Hoover and Perez (2000), which is used by Mayer (2000), Pagan and Veall (2000), and Spanos (2000): "'Data mining' refers to a broad class of activities that have in common a search over different ways to process or package data statistically or econometrically with the purpose of making the final presentation meet certain design criteria." (p. 196). This is a good definition, but there are a number of small issues I would like to change. I will discuss some of its features and discuss them critically to come up with my own definition.

Hoover and Perez describe data mining as a class of activities that are *intended* to meet certain design criteria. This excludes unintended data mining, which is arguably a real option (Denton, 1985; Caudill, 1990; White, 2000; Greene, 2000). I therefore take it to be too narrow. There is some ambiguity in the way the term is used though. Some authors (Spanos, 2000, included) use the term as describing an activity. Others (Denton, 1985 and Caudill, 1980, included) talk about "collective data mining" which may occur even if all the individual researchers abide by the rules of the standard view on econometrics. In order to allow for the latter meaning of the term I propose that data mining does not refer to an activity by itself, but that it refers to the process by which statistical evidence was brought about. If this is purposefully done by one researcher, it is intentional data mining, if not, it is unintentional data mining.

Another part of the definition that may need some clarification is "*certain design criteria*". In economics, the design criterion used is almost always a statistical measure that is based on fit with the data[4]. One particular good way to look at it, is to point out that most criteria in statistics are supposed to tell us something about the chance the evidence came about were the hypothesis false. In case of data mining, model specifications are sought that either have statistical properties reflecting either a small or a large chance. An important term in Hoover and Perez's definition that I will take over is that of "search". If data is selected already for a statistical quality with respect to the data in one form or another, it is a matter of search. Search entails that a number of hypotheses were under consideration, and the one with the best evidential quality was in the end selected.

I can now reformulate Hoover and Perez' definition in order to form my own:

Data mining is a process in which statistical properties of the data are used in the formulation, or selection, of hypotheses that are themselves evaluated by means of related statistical properties of (not necessarily the same) data.

1.3.2. Scope, degrees and weak data mining

I briefly want to make some further clarifications. Recall that some authors worry that the way in which data mining is defined is unhelpful as it characterizes almost any research activity as data mining (Mayer, 2000). After all, it is almost never the case that people know the model beforehand completely and can specify it without making any adjustments based on the data. I did not make a substantial effort in my definition to avoid this. That is because saying that it is data mining if you look at 100 models to select the best one implies that it is also data mining if you do the same with 2 models. It would be strange to let the definition depend on the scale. However, it is important to take into account that there are different degrees of data mining.

Furthermore, another distinction needs to be made. Hendry and Krolzig (2004) cite Gilbert (1986) to make a distinction between weak and strong data mining that can be used to make an important clarification. There is a large difference between the researcher who actively looks through the data to find a good specification about which he can write papers, and the researcher who uses automatic search methods to find specifications they believe are most likely to be true. This distinction is methodologically very relevant and it is categorical. Hendry and Krolzig argue that the cases of automatic data mining are weak forms of data mining as they are not likely to affect the reliability of the data much, while the cases where researcher do the searching themselves with the purpose of finding the best representable result, without much care about how reliable the method is by which he achieved it, is much more likely to results in bad evidence. The latter form is thus strong data mining. The case I discuss in chapter 2 is one of automatic search algorithms applied on a large scale, and is therefore a case of weak data mining. Cases of strong data mining will be discussed too, but we need to keep in mind that there is a difference between these two forms of data mining.

[4] This may be in the form of t-test for particular parameters, or in the form of f-tests for complete models.

1.4. The two main problems of data mining

I already discussed a number of worries related to data mining, but I present two worries that will be the core of my thesis in more detail. These two worries are perhaps not an exhaustive overview of the problems related to data mining, but they are the most important ones.

The first problem is related to the fact that data mining is considered a violation of the rule that one should have a proper hypothesis before one looks for evidence for this hypothesis, and not the other way around. Secondly, an important problem of data mining is that the multiple testing that is involved distorts the standard interpretation of statistics. Both these problems have been cited as the most serious problem of data mining. The problems are discussed in different fields of philosophy. While the problem about the interpretation of test statistics in case of data mining is discussed in the philosophy of statistics, the problem about the double usage of the data is discussed in a different debate in which both general philosophers of science and philosophers of statistics participate. The latter is a general debate about the matter to what extent evidence should be novel, while the latter is a more specific debate about how statistics can be interpreted when independence of the formulation and evaluation of hypotheses cannot be assumed.

I briefly discuss the problems in some more detail, and explain how I treat them in my thesis.

1.4.1. Research without theory

A first major source of concern with data mining is that data mining very often uses the same data to form hypotheses and to use as evidence for these hypotheses. Hence, there is a double usage of the data (Mayo, 2008; DuPlessis, 2009; Hollanders, 2011). This is said to be problematic because it is ad hoc. There are some deep worries in the literature in philosophy of science about ad hoc inferences. A theory that is constructed ad hoc fits with the data per definition, but this fit does not provide any confidence that this theory is in fact true. A simple example is that of curve fitting. If a person wants to examine how the returns on a certain stock behave over time, one can always find a relationship that fits with the data perfectly (e.g. Lo and MacKinley, 1990). However, because there are infinitely many alternatives that do the same, the extrapolation of this particular formula is not considered very reliable. Examples like this make philosophers doubt that any post hoc constructed theory can be said to be supported by the data. However, there are fierce debates about this.

Worrall (2010) defends the view that double counting is bad scientific practice. For him evidence needs to be novel in order to qualify as evidence. The reason for this he traces back to Duhem. However, both classical and Bayesian philosophers of statistics have objected against this criterion of evidence.

In chapter 3 I discuss these arguments in much more detail and examine them in the context of data mining. An important and interesting fact for our purpose is that not all instances of data mining are also instances of double counting. This allows us to analyse the harm done by double counting in isolation.

1.4.2 Statistical reasoning and pretest bias

The bias in statistical interpretation of results due to data mining is perhaps one of the most worrying aspects of data mining. The problem of the pretest bias, or multiple testing, is most easily explained in, but not restricted to, the context of classical statistics. The central idea of this method of inference is that data that are highly unlikely to be observed given a hypothesis are an indication that the hypothesized model, H, is incorrect. This conditional probability is called P(E|H). The estimated value of P(E|H) is called the p-value. The p-value is an estimate of that chance[5]. If it is 5%, it means that in 5% of the cases we would expect a deviation to be observed as large as or larger than the one that was observed. A low value is unexpected if the hypothesis is true. For the classical statistician, this is evidence against the hypothesis. Ronald Fisher (1956) described it as follows: If one observed a unexpectedly large discrepancy from the expected value according to the null hypothesis (H) *''Either an exceptionally rare chance has occurred, or the theory of random distribution* [i.e. H] *is not true*". (p. 39). The rationale behind this interpretation is that unlikely events are unexpected. If the hypothesis is in fact true, we would not expect to observe a low p-value. So, if we do observe a low p-value, this is an indication H is false. The probability that is considered to be "exceptionally rare" is commonly taken to be 5%[6], this is called the alpha, α, of a test, and represents the false-negative rate: the probability that one rejects the hypothesis, while the hypothesis is in fact true. This probability is called the probability of a type I error. The propensity of a test to make type I errors is called the size of the test. Hence, the smaller the size of the test, the less type I errors are made. Type II errors are another mistake statistical tests can make: not rejecting a false hypothesis. If a certain coefficient that is indeed different from zero is tested to be insignificant, a type II error has occurred. The propensity of a test to make type II errors is called the power of a test. The higher the power of the test, the less likely the test is to make type II errors.

Now, the problem that data mining poses is that not just one test is run, but a large set of them. And it is a simple consequence of a probability interpretation of samples that rare outcomes are simply expected to happen once in a while. Even if all the null hypotheses that are tested are in fact true, one is expected to find a lot of rejections if a lot of tests are run. If a very low p-value is found, because it was sought for, it is much more likely due to statistical randomness than in case evidence was found for a hypothesis that was formulated independently of the data. Hence, searching in the data and neglecting this search is misleading, and the statistical results will be biased. This bias is called the pretest bias, because the large number of tests conducted before one significant test was found make it likely that something significant was found even if there is nothing genuine to be found. This bias is a major source of concern for results found by means of data mining (Lovell, 1983). Mining the data almost necessarily will provide some significant p-values, but what a researcher is to conclude from it is not always clear, because of this bias.

[5] Technically, classical p-values measure something different (Keuzenkamp and Magnus, 1995): the chance that a larger discrepancy from the expected value is found than the observed discrepancy, were they hypothesis correct. This has a different meaning than P(E|H).

[6] Keuzenkamp and Magnus write: "if economists have natural constants, then the most well-known is .05" (p.16).

In order to illustrate that the problem of pretest bias is very real, consider the following example from Austin et al. (2006). Austin et al. conducted a large experiment in the Canadian providence of Ontario. Data on all 10 million inhabitants in 2000 was available. They examined whether there were any astrological signs that would make it more likely for people to suffer from a certain disease. They found that there were 24 statistically significant associations of diseases with astrological signs: two diseases for each sign. For instance, people with the astrological sign Leo were significantly more likely to suffer from gastrointestinal haemorrhage (P = .0041), and the astrological sign Sagittarius was associated with humerus fractures (P = .0458). The data set was then split up in two cohorts. One was used to find these associations; another was used to validate the result. While in the second cohort most of the associations indeed disappeared, the association of Leo and gastrointestinal haemorrhage and the association of Sagittarian with humerus fractures remained significant.

The authors themselves wanted to point out the dangers of what they call multiple testing. They show that if proper corrections for multiple testing were used, the found significance would have disappeared. This is both the case for their first and their second cohort. Moreover, they opt for guidance of theory in empirical investigation. The extent to which these warnings are justified will be discussed in chapter 4 and 5.

1.5. A look ahead

Data mining is methodologically ambiguous and surrounded by many controversial debates. While it is considered a sin to many, some endorse it as a research philosophy. Reducing the normative and methodological ambiguity about this topic is a key aim in what follows. In the following chapters I discuss data mining and its related problems in more detail. As a look ahead, I present a short description of what will be argued in the different chapters of this thesis.

1) There are instances in economic research in which there is little theory available and one has to study the data in order to form good hypotheses about the matter of interest. (chapter 2)

2) The double counting objection against data mining does not seem to be a good reason, in principle, to dismiss all cases of data mining. (chapter 3)

3) The pretest bias objection against data mining is a convincing reason to condemn data mining. But, it applies to a very large share of cases. Pretest bias can be avoided by adjusting the test statistics in proportion to the degree of search, but the degree of search is not something we can learn practically, or understand conceptually in economics. Consequently virtually every empirical research in economics is subject to the pretest bias objection. (Chapter 4)

4) Alternative interpretations exist that allow us to still learn from test statistics, without interpreting them in the classical manner. There is good evidence that some automatic search

procedures (in particular *Gets*) are reliable procedures under certain conditions. But how these methods perform under real conditions remains a black box. (chapter 5)

5) P-values should not be interpreted inferentially in the case of empirical research in economics. (chapter 4, 5 and 6)

Chapter 2: A Quest for Truth in an Open-Ended Field of Economics

"[D]ata mining is the worst possible way to use data to learn about important economic issues, except for all the others ways."

Pagan and Veall (2000, p. 216)

2.1. Introduction

In chapter 1 I discussed the *Condemnation Thesis* (**CT**): data mining is bad practice and should be avoided. In order to do so hypotheses should be formulated independently of the data. Without going into a conceptual discussion on the difficulties with this view (that will follow in chapter 3 and 4), we can question the fruitfulness of this rule on the basis of practical criteria. In this chapter I want to discuss one instance in which the rule is certainly not fruitful. This instance is growth economics.

Growth economics is widely discussed in the literature on data mining (e.g. Hoover and Perez, 2004; Hendry and Krolzig, 2004; Du Plessis, 2009). In fact, Sala-i-Martin's (1997) paper *I Just Ran Two Million Regressions* was a paper in the field of growth economics. This is no coincidence. There are a number of problems that arise in the field of growth economics. These problems have triggered the use of data mining methods. The aim of the chapter is twofold. The first one one is to present a case in which maintaining **CT** is unfruitful on practical grounds. The second aim of the paper is to provide some insights in the kind of data mining methods that have been developed.

The degree to which the data mining methods that have been applied in the field of growth economics are warranted will not be dealt with in this chapter. The argument made in here is merely that at least one reason why researchers went to data mining methods in order to find answers to their questions is that there was no other promising way to find answers. This chapter can be read as a plea for pragmatism about data mining. Not necessarily because data mining is not as bad as may be thought, but rather because for some questions data mining is the only way to learn.

2.2. Growth theory: an open-ended field of research

The main question the field of growth economics deals with is: What are the main determinants of long-term growth? Or, put slightly differently: Why are some countries rich and others poor? Until the 1990's, the main theoretical framework to answer the question was the Solow growth model. Robert Solow's model (1956) was built upon a number of empirical regularities that were later described as the Kaldor stylized facts (Kaldor, 1957). Kaldor observed that the capital intensity (amount of capital per worker) and efficiency (amount of output per worker) keep increasing, while the capital output ratio (the amount of capital per output) is roughly constant. Solow's model depicts the whole economy as a production function where the inputs are technology, labour and capital. The output is aggregate

production. The labour input is proportional to the population, while its effectiveness to produce is determined by the available capital and technology. Solow's simple model assumed that the output of people increased with decreasing returns. That is, the first person in the economy can increase production by a lot, while the one hundredth person in the economy will add much less to the production. The same accounts for capital. The first machine increases efficiency much more than the one hundredth machine. Because people save and invest part of their income, the amount of capital increases. However, machines sometimes break down, and lose their value for the production process. This is called depreciation. As the marginal effect of machines decreases, a point will be reached at which the new investment will be equal to the depreciation of the old machines. In other words, all the new investment will be offset by machines breaking down, and rather than investment resulting in more capacity, it is used to maintain the standard of capacity. At this point, an equilibrium is reached. The only reason why economies still grow, at this point, is due to advances in technology.

While the Solow model was the accepted framework to analyse economic growth for a long time, there were two sources of dissatisfaction about it among economists (Romer, 1994). The first is that the model predicted that the economies of different countries would converge in the long-run. In other words, countries with lower outputs per capita have a higher growth potential. This is called the Convergence Hypothesis. The Convergence Hypothesis is a consequence of the fact that the model assumes decreasing returns to capital: when there is not much capital available in the country, a little bit of extra capital will make a large difference to output. As developing countries are generally not capital intense, it is expected that small investments will have large impacts. This, however, is not at all what is observed in the data. Poor countries did not seem to "catch up" on the developed countries particularly fast, while the developed countries kept developing.

A second reason the Solow growth model was considered unsatisfactory is that once in equilibrium the explanatory power of the model to explain further economic growth is very limited (Romer, 1994). In equilibrium, per capita economic growth is determined by technological process only, which is *exogenous*. That means that it is not itself explained in the model.

These two criticisms led, at the end of the 1980's, to a number of alternative theories to explain growth. These models were called Endogenous Growth Theories or New Growth Theory because they tried to find ways in which growth *would* be explained by factors within the model. That is, by *endogenous* factors. Seminal papers are Romer (1986) and Lucas (1988). Romer (1986), for instance, builds a model which tries to explain technological growth by means of the percentage of the labour force devoted to technological progress. Lucas (1988) assesses the possibility that returns to (human) capital are stable rather than diminishing as in the Solow growth model. This explains why equilibrium is not reached, why there is no convergence and why human capital appears to be so important in explaining economic growth.

These New Growth Theories implied a paradigmatic shift in thinking about growth. Given the problems Solow's model suffered from, these theories were in great demand. However, with this new way of thinking new problems arrived. While Solow's growth model may have been unsatisfying for economists given its limited ability to explain growth, its

simplicity had many advantages. The New Growth Theories opened up the field for explanations beyond the standard neoclassical cookbook. A crucial feature of New Growth Theory is that different theories propose a number of possible causal factors of growth, but do not exclude any factors (Ulasan, 2011). Brock and Durlauf (2001) refer to this feature as the open-endedness of the growth theories. Concretely, this boils down to the fact that there have been a large number of variables selected on theoretical or empirical grounds, while we have no grounds to say which are more plausible than others. Sala-i-Martin (1997) considered 62 variables proposed by different theories, while an overview by Durlauf et al. (2005) finds 145 different variables that have been proposed in growth studies and 43 distinct growth theories in which they appear. The question which of these 145 variables really explain growth remains difficult to answer. As Wacziarg (2002) puts it: a large variety of variables explaining growth tend to have their "15 minutes of fame" (p.907). It seems unlikely that economists will be able to select the best theory of growth on purely theoretical reasons. It therefore becomes inevitable that the decision which model is the best model for growth requires empirical investigation.

2.3. Growth empirics

The relationship between the empirical research and theory is that theoretical models have statistical counterparts. Theories propose causal factors that explain growth, statistical models test to what extent operationalized theoretical models fit with the data. Statistical measures of fit are generally based on measures of conditional correlation, that is correlation , while keeping other variables in the model constant. For example, a theory of growth that emphasizes human capital as an explanatory factor of growth can be tested by a regression model in which a measure of human capital, such as number of years of education, is correlated to growth, in which a number of important covariates, such as investment, technological progress and government are put in the model too, to make sure that the correlations between human capital and growth are not due to these factors. They are "kept constant" in a statistical sense[7]. In what follows, I refer to both simple correlations and conditional correlations in statistical models as statistical relationships. If such statistical relationships are not genuine they are spurious. However, spuriousness can mean two things. A spurious statistical relationship may mean that there is a genuine correlation that is due to a common cause, or that that the correlation itself is not genuine, but due to chance. I use it in the latter sense.

If good measures of theoretical constructs are found, a statistical fit with the data is an implication of a good model of growth. However, this does not work the other way around. A good statistical fit does not imply that the theoretical model is likely correct (see the discussion on Haavelmo, 1944, in chapter 1). Nevertheless, empirical research on the topic has generally attempted to consider which variables among the large number of those proposed by different endogenous growth theories are correlated to growth (given the relevant control covariates). Sorting out which variables correlate to growth, given the right number of covariates, helps to reduce the set of admissible theoretical models, even though it might not

[7] Not in a causal sense, statistical models can generally not provide causal ceteris paribus conditions.

yet, by itself, teach us what the true model of growth is[8]. A typical representation of a statistical model is the following regression. This is called a Barro regression, after Barro (1991):

$$Y = \beta_0 + \beta_i X + \beta_j Z + u \quad (1)$$

Where Y is the level of long-run growth, β_0 is a constant, X is a vector, or a list, of variables that are always included in the regression because their relationship with growth is considered to be well-established. These variables serve as covariates. β_i is a vector of parameters to be estimated that capture the effect of the variables in X. Z is a vector of variables whose effect on economic growth are of interest to the researcher, and β_j are the corresponding parameters to be estimated. X typically includes three or four variables (e.g in Levine and Renelt, 1992 and Sala-i-Martin, 1997 respectively). Typically they include the level of output in the starting period, and the level of private investment. Z can include any subset of the 145 variables that have been suggested on theoretical grounds. Needless to say, regression (1) can take an almost endless amount of concrete forms. Ley and Steel (1999) for instance, test all possible models that can be built with 41 variables and end up with over 2 trillion regressions; while Hendry and Krolzig (2004) find that there are over a million trillion different possible regression models that can be made with the available variables proposed by all the different theories.

A simple solution to the problem would be to say that all these factors should be regarded as important factors to explain growth. However, not only would this result in a very complex and unparsimonious growth model; there is a further problem. Such a model is very difficult to estimate empirically, and therefore, it is quite useless as a method to find out which of the proposed factors are in fact important and which ones are not. There are two reasons for this. Firstly, there is a serious problem of *multi-collinearity* of regressors. Secondly, a serious case of data poverty plagues this field of research. The number of countries in the world is limited, which means that there is only a limited number of long-term growth differentials to be observed. I will discuss these two issues briefly.

Firstly, there is the issue of the multi-collinearity of regressors (henceforth **MC**): one variable may be a cause of growth and a statistically significant factor in a large number of regressions, but as soon as a highly correlated covariate is added to the regression model, the statistical significance will evaporate (collinearity means that two regressors are strongly correlated). It is a crucial assumption of the most common regression model estimation method, the method of the ordinary least squares, that the regressors in a regression model are uncorrelated in order for the estimates to be unbiased (see econometrics text books, such as Verbeek, 2008, or Wooldrige, 2009). If this is not the case, the estimated parameter will be biased. Intuitively this is because these two variables both partly explain the same variation. Consequently, it becomes more difficult to attribute which part of the variance in the target

[8] I accept that "true determinant" of growth is a very unclear term. This formulation however, is not my own. It appears, for instance, in Sala-i-Martin (1997: in scare quotes), Hoover and Perez (2004; who provide a critical discussion) and Ulasan (2011). I therefore continue to use the terminology. True model, or true determinant of growth, can be interpreted causally, such that if a determinant (in a model) is in fact a causal factor. However, I believe that it is often intended differently: a true determinant is in fact associated with the variable of interest in the population. The cause may be spurious, but the correlation is genuine.

variable is explained by which explanatory variable. An example in the growth literature is the "Asian dummy", a variable that is 1 for all Asian countries and zero otherwise, which is highly collinear with the "part of the population that is Confucian". Both regressors are significant when the other is not included in the regression, however, when both are included, only the Confucian dummy is significant. Hence, highly collinear regressors are unlikely to be estimated correctly. At the same time, due to **MC**, it very likely that spurious statistical relationships are found when instead of the true causal factor of growth, a variable that is highly collinear is added in the model. In short, it is very difficult to identify the true importance of the variables that explain growth due to correlations that exist between them. Consequently, a model with 145 regressors is not only unparsimonious, but is also likely to misrepresent the true effect of variables that have correlated covariates.

Practically, one of the consequences of **MC** is that a regressor may be highly significant in one statistical model, but insignificant in another. Sala-i-Martin (1997) puts it as follows: "If one starts running regressions combining the various variables, variable x_1 will soon be found to be significant when the regression includes variables x_2 and x_3, but it becomes non-significant when x_4 is included." (p. 178). Evidently, this is a very undesirable feature of the data when one is interested in which of all the causal factors cooked up by theory and empirical research are the true ones.

Secondly, a problem arises with the available information. Growth economics is not focused on the causes of growth next year, but on the causes of growth in the long-run. A general way to measure this in the empirical literature following Barro (1991) is that for all countries the growth differential from 1960 until today is calculated as the growth variable of interest. Consequently, for all countries for which data is available there is only one data point. Some have proposed panel data techniques (e.g. Islam, 1995). A problem with this is that it does not capture the long-run, but rather the short-run correlates of growth. Many of the data sets that are used contain 119 variables in total (e.g. Sala-i-Martin, 1997; Levine and Renelt, 1992; Hoover and Perez, 2004). But data on all variables is not available for all countries. The more variables are included in the research, the less countries can be used. Sala-i-Martin (2004) used 88 countries. The number of observations however, never exceeds 120, and as a consequence the degrees of freedom for estimating a large model are never sufficient if one is interested in 145 possible explanatory variables.

The theory available is thus not very useful for finding a promising model to explain growth. There are many possible variables, and one large model that contains all of them is not very helpful. A selection of these variables needs to be made. But this cannot really be done by means of the classical method in which one uses statistics only to reject theoretically motivated hypotheses. Theory teaches us only that there are millions of trillions of possible models, and we do not know a priori which ones are more likely than others. This problem has been aptly referred to as a problem of model uncertainty (Brock and Durlauf, 2001; Ulasan, 2011). In order for statistical methods to help us, we need to be able to use them to reduce the total amount of possible models explaining growth. This need has inspired research that applies methods to learn from the data. These methods deviate from the **CT** in that they effectively try to use statistical criteria in the data to form specific hypotheses about the phenomena the data describe.

2.4. Three approaches to model uncertainty

There are roughly three general approaches to dealing with model uncertainty. I discuss these methods and the way they have been applied to growth economics.

2.4.1. Leamer's Extreme Bounds Analysis and Levine and Renelt (1992)

One argument made in Leamer's (1978) book is that it is necessary for applied econometricians to make decisions for which there is no theory available to help them. The key question that Leamer addresses is how to do this in a way that does not make the statistical inference unreliable. One of his suggestions (1983; 1985) is the Extreme Bounds Analysis (henceforth **EBA**). The idea behind **EBA** is that it does not inspire confidence in the usefulness of an explanatory variable if it is statistically significant in some regressions, while it is insignificant in others. Or, even worse, if it has positive sign in some regressions, and negative one in others (or vice versa). Doing an **EBA** on an explanatory variable means that one runs a large number of regressions with the variable of interest and a variety of relevant variables that are expected to explain the dependent variable (growth in this case). And, when that variable is significant and has the same sign in all regressions, it is said to be robust. If not, it is said to be fragile.

The first researchers who attempted to deal with the problem of model uncertainty of the growth literature in a systematic and empirical way were Ross Levine and David Renelt (1992) and they used Leamer's **EBA**. The basis for their research is regression (2), a variation of (1):

$$Y = \beta_0 + \beta_i \mathbf{X} + \beta_1 M + \beta_j \mathbf{Z} + u \qquad (2)$$

where M is the variable of interest. They considered 4 variables in their **X** vector, that would be in all regressions, and considered a larger number of variables of interest, M. **Z** is a vector containing up to three relevant variables from the total set of variables considered. For every variable of interest, three variables were selected for **Z** such that they did not capture the same phenomenon as the variable of interest M. This resulted in seven regressions for every variable of interest. Levine and Renelt then applied Leamer's **EBA**. They found only 1 variable that was a robust determinant of growth in their research: average share of investment.

The reason this is data mining is because **EBA** is based on discovering robustness of models with respect to a certain data set, while robustness is based on fit with the data. Robustness is used to arrive at conclusions about which variables belong in the true model, while robustness is also taken as a good reason to believe the relevance of one particular variable to explain the other. Hence it is using a measure of statistical fit (in a variety of statistical models) to both construct and evaluate statistical hypotheses.

2.4.2. Bayesian Model averaging and Sala-i-Martin (1997)

Xavier Sala-i-Martin (1997) argues that Levine and Renelt (1992) took the right step in going from testing single theories to testing the large number of proposed regression models together. However, he believes that the **EBA** as they have conducted it, is too strict. Due to the **MC** problem, he argues, it is almost impossible to find a variable that is robust in all regressions, even if it is a true cause of growth. He therefore proposes a different method to go about the problem of the open-endedness of growth theories. First of all, he considers a great deal more regressions than Levine and Renelt (hence the title of his paper: *I Just Ran Two Million Regressions*). However, he does not use the same robustness analysis as Levine an Renelt. Sala-i-Martin proposes to use regression (2), with three variables in **X** and exactly three variables in **Z**, but to alternate these three variables, such that all possible combinations of the 58 variables (62 – 3 **X** variables – the M variable) are estimated. This results in over 30,000 regressions per variable in the data set, arriving at a total of 2 million regressions in total. What is different from Levine and Renelt is that he then proposes a weighted averaging technique to aggregate these results. For all estimated regressions, a likelihood is calculated: a probability of how likely it was that the data was brought about, were the regression the true regression. The coefficients of all tests are aggregated by taking their averages weighted by their relative likelihoods. The same is done with the estimated confidence bounds (or standard errors). This then, results in an overall p-value. If that p-value falls below the conventional 5% level, the variable is said to be a robust correlate of growth, if it is higher than this, the variable is said to be a fragile correlate of economic growth.

Following the publication of Sala-i-Martin's (1997) article, a number of follow up articles have been published on the topic. A number of approaches have arisen. Following Sala-i-Martin (1997) a literature on ways to average model estimates such that they present a reliable estimate of the true effect has emerged. These papers assume that every regression coefficient provides a certain level of confidence we should have in the relationship between the variables in the regression and economic growth. The methods have to find some way in which the confidence can be aggregated, which can only be done in a Bayesian framework. The approaches fall in roughly two categories (Ulasan, 2011): firstly, Bayesian Model Averaging (**BMA**; Ley and Steel, 1999; Fernandez, Ley and Steel, 2001) and, secondly, Bayesian Averaging of Classical Estimates (BACE; Sala-i-Martin et al, 2004). The general approach of these methods is similar in spirit to Sala-i-Martin's (1997a) approach, but generally includes some method to compare models that are of different sizes, while Sala-i-Martin's approach only allowed for comparison of regressions with the same amount of regressors (in his case seven).

There is a subtle difference in interpretation between the classical approach to statistical inference and the Bayesian approach if it comes to the interpretations of the probability that plays an important role in case of data mining. Classical econometricians only estimate conditional probabilities: the probability that the data is observed, were the hypothesis under examination true[9]. The inference made on this basis is that if this probability is very low, it is an indication that the hypothesis is false. However, the latter probability is

[9] That is, a version of this probability. It in fact estimates the chance that something even more deviant from the expected value is observed, which is not exactly the same as the conditional probability (see note 5).

not actually calculated. The reason for this is that the calculation of this probability requires more information than can be observed under the assumptions that frequentist statisticians want to make. Namely, information is needed about the probability that the hypothesis is true before the observation of any evidence. Bayesian statisticians call this the prior (probability). Classical econometricians shy away from this type of inference, as there is no objective way to observe prior probabilities in the world, and they want to avoid subjectivism in inference. They only follow the reasoning based on the conditional probability.

In case of data mining this causes a problem. Data mining in the context of cross-country growth regressions involves running a large number of regressions to infer which variables generally work well, and which ones do not. However, if so many different models are estimated, one is bound to observe an insignificant coefficient even for the strongest effect due to MC. In the classical framework there is no other way to aggregate these test than to say that they are either rejected or not. Averaging p-values would assume that the tested hypotheses are comparable in likelihood, which is something that is deliberately not assumed in the classical framework. However, the methodological approach is completely silent on what to do with the question what to believe when there are regressions in which a variable is significant, and some regressions in which it is not. Consequently, the aggregation of different coefficients becomes inevitably Bayesian. Sala-i-Martin's (1997) approach, for instance, is implicitly Bayesian in spirit, because they apply such an aggregation method (Sala-i-Martin et al., 2004).

The field of Bayesian model averaging is currently one of the mainstream methods to deal with model uncertainty in growth economics (e.g. Magnus et al., 2010), and is still developing. Recently, a number of authors have argued for augmentations to the **BMA** in order to allow for non-linear models and heterogeneity of parameters for different countries (e.g. Cueresma and Doppelhofer, 2007; Salimans, 2012).

2.4.3. The general-to-specific approach and Hoover and Perez (2004)

An alternative method to doing econometrics in general, and data mining in particular, is the so-called LSE approach popularized by David Hendry. Key in Hendry's methodology is the concept of *reduction* (Campos et al, 2005). One starts with a model containing all the explanatory information that is available, and then the redundant information is "scraped off". As such, the method reduces a large model to a smaller one. This can go on until no further valid reductions can be applied. This philosophy of inference was developed by David Hendry and co-authors in the 1980's and the most important papers in which this approach is developed are collected in Hendry (2000). In the late 1990's, Hoover and Perez developed a way to implement the approach systematically in a computer programme that has become the General-to-Specific algorithm (henceforth *Gets*; Hoover, 1995; Hoover and Perez, 1999; 2000; 2004). The approach was then further developed by Hendry and Krolzig (1999; 2004; Krolzig and Hendry, 2001). The algorithm starts with a model containing all the considered determinants of the dependent variable. Hendry refers to this as the general unrestricted model (GUM; Hendry and Krolzig, 2004). The model is reduced by removing the insignificant regressors, step by step. The re-estimation of this model will repeatedly result in reductions of

the general model until no more insignificant t-statistics appear in the model: this, then, is the final model. This is the general strategy of the *Gets* methodology. However, a problem with this method considered by Hoover and Perez (1999; 2004) and Hendry and Krolzig (2004) is that due to MC there may be a considerable path-dependency in what model is selected in the end based on which insignificant variables are removed first. Hence, they extend Hendry's original method by allowing for different selection paths. Other important additions to the *Gets* methodology are the fact that a number of diagnostic tests are included in the selection path. For instance, a test for heterogeneity is added to the algorithm by Hoover and Perez (2004).

Hoover and Perez (2004) and Hendry and Krolzig's (2004) were both responses to the selection methods that were used in the growth economics debate. In particular they were responses to Sala-i-Martin (1997) and Levine and Renelt (1992). Both groups of authors argue against **EBA** as used by Levine and Renelt and against **BMA** as used by Sala-i-Martin (1997) and Fernandez, Ley and Steel (2001). The fact that these methods solely rely on the robustness of variables is found unsatisfactory. Supporters of *Gets* emphasize the importance of diagnostics testing and evaluating econometrics models on their realism rather than on the robustness of the variables in the model.

The *Gets* approach is very different from the **EBA** and **BMA**/**BACE** approaches discussed above. However, it is still a form of data mining in our definition. The *Gets* algorithm selects from the large possibility of models one that is a valid reduction from the general model and still explains all aspects of the data that the general model described. The model that roles out of this procedure is taken to be the most plausible model (e.g. Hoover and Perez, 2004). The criteria to reduce the GUM are all based on statistical measures of the variables that are removed and of the resulting model. The statistical criteria determine the final model, while the fact that the it was brought about by the method is taken as evidence for the model. Hendry and Krolzig (2004) pronounce in their title *"We ran one regression."* The *Gets* methodology does use a lot of significance tests to reduce the GUM, and this title says very little about the amount of statistical testing that is used to arrive at the final model (this point is also made by Keuzenkamp, 1995). The reduction from the GUM to the final model is thus based exclusively on statistical criteria.

2.4.4 Some results

In table 1 below, the findings of a number of papers discussed above are summarized. Firstly, it is striking that there is a large variety in the number of coefficients that are found to convincingly determine growth according to the different methods. We can see that the **EBA** (column 1) is a particularly stringent method: it only selects one variable, while both the modified **EBA** (column 2) and the BACE select a very large amount of variables. **BMA** only finds four and *Gets* finds five variables. In column 6 a more recent paper is summarized. Tim Salimans (2012) developed a sophisticated method to allow for non-linear models in **BMA**. It concludes that much less variables are robust when in his method compared to the standard, linear, **BMA**.

Another striking feature is the small overlap in selected variables by different methods. In all fairness, all papers conclude that at least one investment variable is relevant (either investment share of GDP, non-equipment investment, or investment price). Moreover, 4 out of the 6 methods conclude Fraction Confucian is an important determinant of growth. Only one paper concludes that Years of Schooling is important, and only three out of the 6 papers conclude that the Years Open Economy is important.

Table 1: Results of the discussed papers

	(1) Levine and Renelt (1992)	(2) Sala-i-Martin (1997)	(3) Fernandez, Ley and Steel (2001)	(4) Sala-i-Martin, Doppelhofer and Miller (2004)	(5) Hoover and Perez (2004)	(6) Salimans (2012)
Method used	Extreme Bounds Analysis	Modified Extreme Bounds Analysis	Bayesian Model Averaging	Bayesian Avergaing of Classicial Estimates	General-to-Specific	**BMA** allowing non-linearities
1	Investment share of GDP	Equipment investment	GDP in 1960	East Asia dummy	Revolutions and Coups (-)	Investment price (-)
2		Number of years open economy	Fraction Confucian	Primary schooling 1960	Fraction Protestant (-)	East Asia Dummy
3		Fraction Confucion	Life expectancy	Investment Price (-)	Fraction Confucian	GDP in 1960 (log)
4		Rule of Law	Equipment investment	GDP 1960 (log) (-)	Equipment investment	Air distance to big cities
5		Fraction Muslim		Fraction of Tropical area (-)	Number of years as an open economy	Fraction GDP mining
6		Political rights (-)		Population density coastal 1960's		
7		Latin America dummy (-)		Malaria prevalence 1960's (-)		
8		Civil Liberties (-)		Life expectancy in 1960		

9	Fraction GDP Mining	Fraction Confucian
10	SD Black-Market premium (-)	Africa dummy (-)
11	Primary exports in 1970 (-)	Latin America Dummy (-)
12	Degree of Capitalism	Fraction GDP in Mining
13	War Dummy (-)	Spanish Colony (-)
14	Non-equipment investment	Years open economy
15	Absolute latitude	Fraction Muslim
16	Exchange rate distortions (-)	Fraction Buddhist
17	Fraction protestant (-)	Ethnolinguistic fractionalization (-)
18	Fraction Buddhist	Government consumption share 1960's (-)
19	Fraction Catholic (-)	
20	Spanish Colony	

It is important to note that all data mining methods discussed here are all based on finding the best performing correlational structure. These methods do not include a causal interpretation of these correlations. In line with Koopmans (1947), we can say that simply more theory is needed for this. However, most defenders of these methods do believe that their method is a fruitful way to learn something about the causal structure that brought about the data, in the sense that their methods are aimed at reducing the number of possible variables that may be important to growth. Establishing genuine statistical relationships can fruitfully lead to causal knowledge by applying the Reichenbach principle (Reiss, 2008): a genuine correlation is either due to a common cause, or due to a causal relationship that exists between the two correlating variables. By providing more knowledge about the correlational

structure of the data, many possible causal models may be excluded. While Leamer objects to the notion of a true model to describe economic phenomena, he does argue that robustness is a sign of usefulness to explain a phenomenon. The *Gets* approach is particularly ambitious. It tries to add a number of additional criteria to mere correlational analysis to put further constraints on what one would expect to find in the data if the data was brought about by a fixed causal structures. Tests to detect hidden patterns in the data that are not explained by the model are an example of this. In the end, though, it requires background theory to explain a correlational structure causally. This is acknowledged by most authors, and no one claims otherwise.

2.5. Data mining and alternative ways to deal with open-endedness

The discussion above summarizes the main ways in which the cross-country growth regressions have been analysed. All three qualify as methods of data mining according to our definition. In chapter 1 we have examined some problems of data mining that will be discussed in much more detail in the following chapter. These problems explain why data mining was considered bad practice in the profession (Leamer, 1978; Kennedy, 2002). However, before we would dismiss any methodology for breaching rules, we may ask: What are the alternatives?

There are actually a number of other methods that are used to study growth. Joshua Angrist and Jorn-Steffen Pischke (2010) wrote an article in which they discussed the way in which growth economics has developed over the past decades as a response to critiques voiced by Leamer in particular and other economists of the 1980's in general (e.g. Sims, 1980). They argue that approaches based on Leamer's Extreme Bounds Analysis, in which they include Sala-i-Martin's (1997) paper, has disappointed in its delivered results. The really promising results, they argue, come from different movements in econometrics, namely those focusing on instrumental variable approaches and natural experiments. In case of growth theory they mention an influential paper by Acemoglu et al. (2001) who investigate the effect of institutions on growth by gathering historical data on settler mortality rates to use as an instrument. Another important direction of research that has been taken to research economic growth is the randomized controlled trials movement that has tried to explain growth developments on the micro level (e.g. Banerjee and Duflo, 2011). Both approaches have been celebrated. Do they make cross-country growth data analysis redundant?

From the turn of the millennium onwards a number of authors started to use instrumental variable techniques to learn about causal relationships in growth economics (e.g. Frankel and Romer, 1999; Acemoglu et al., 2001). And since then, the application of these methods has grown tremendously in the growth economics literature. Similarly, natural experiments have also become more popular, and resulted in very useful insights. For example, Feyrer (2009) uses data around the closing of the Suez Canal in 1967 as a natural experiment to determine the causal effect of sea distance on trade, and thereby, of trade on economic growth.

The great advantage of these methods is that, if you believe the background assumptions that are required for inference on the basis of these methods, which they often

argue for at length, it follows directly that the observed statistical relationship is causal, which generally is incredibly valuable information (see, Reiss, 2008, ch. 7 for a methodological appraisal of these methods). Angrist and Pischke argue that the application of these methods has been very successful and that they are superior to the research inspired by Leamer's **EBA**.

While these are indeed powerful methods when all the assumptions that they require are defensible[10], an important caveat of these approaches is that the conditions necessary for these assumptions are rare. Moreover, their scope is often limited. I do not want to question the usefulness of following this line of research, but merely its limitations. Instrumental variable approaches can teach us something about the causal effect one particular variable may have on growth, but not if it is more important than other variables. In order to do this, one needs to have good instrumental variables for all the variables one wants to compare it too. However, good instruments are rare, and finding good instruments would not be possible for all variables proposed by theory. In this respect, Acemoglu et al. (2001) is a telling example. They argue convincingly, that colonial history is a good instrument for institutions today, and can show that there is a causal effect of institutions on observed growth differentials observed. However, they acknowledge this specific cause of growth differentials (colonially induced institutional differences) does not tell us much about the mechanisms working on economic growth today.

Similarly, randomized controlled trials (RCT's) have become widely used as a method to study economic growth too (e.g. Banerjee and Duflo, 2011). And serve, to some extent, as an alternative to the cross-country growth regressions. Like instrumental variable techniques and natural experiments, RCT's are convincing because if conducted correctly they can prove a causal relationship (Cartwright, 2007). An important disadvantage is that their scope is narrow and that they cannot answer the same kind of questions those who study cross-country growth regressions seek to find (Rodrik, 2009). Even more so than the case of instrumental variables and natural experiments, RCT's are not capable to find answers to the same questions researchers of the cross-country growth regression set out to answer. RCT's can only be conducted on a micro-scale. It is simply impossible to perform an RCT between countries. RCT's may teach us whether small micro treatments work well, or do not work well. However, this is a very different kind of question than the question What are the most important factors are that drive cross-country growth differentials? RCT's are not a very useful method to answer this question. I cannot even start to imagine how RCT's might be used to find an answer to the question why the Asian Tigers developed, and the African countries did not.

In short, besides the analysis of cross-country growth regressions, alternative methods to learn about economic growth have emerged. What is important is that there is important information in the cross-country growth regressions that cannot be fully extracted from it by means of the alternative methods. The alternative methods are thus no close substitutes. Because these methods provide causal knowledge they may have advantages over the cross-country growth regressions in some sense, but the scope of these methods does not stretch to the kind of information contained in cross-country growth data. Cross-country growth

[10] I have to note that Leamer (2010) and Sims (2010) disagree. They both think Angrist and Pischke overstate their case.

regressions are thus by no means redundant in light of the other approaches. In order to answer the question "What are the most important factors related to growth?" there is no other way than to look at all the possible factors and compare them. And therefore we need to analyse the cross-country growth data.

In light of the open-endedness of growth theory, the only efficient way to analyse cross-country growth data is to mine the data. Consequently, we need to mine data in order to learn about cross-country growth differences. There appears to be a stark contrast between this claim and the way I discussed the data mining in chapter 1. In chapter 1 I argued that it is the most common view among economists to condemn data mining, even though everyone knows that it does happen in practice. If this is true, it is at least hard to imagine why growth economists embrace this methodology with open arms. They even pronounce their "sins" proudly in their titles. How can we understand this practice?

The short answer to this question is that there are no real alternatives. It seems that if we want to learn something about this important field of study, there is no other way than via the path of data mining. The methodological problems may be severe, but at the same time, the stakes are high. Learning about the process that brought about the differences in wealth across nations is of crucial importance to human welfare. In particular to those at the bottom of it all. Analysing cross-country growth data is one important way to learn about the correlations existing between social and economic variables related to this question, which is an important step to answer the question. Even if one takes all instances of data mining to be problematic methodologically, it may be the only thing we have to learn about some aspects of this question.

A plea for pragmatism has been made by a number of authors (e.g. Caudill, 1990; Pagan and Veall, 2000). The case of growth economics nicely illustrates why this appears to be the only reasonable position to take on this subject. Even without a methodological discussion on data mining, we can understand why growth economists turned to these methods. It is in this vain that I opened my essay with a quote from Pagan and Veall, who defend the pragmatist perspective: "(…) data mining is the worst possible way to use data to learn about important economic issues, except for all the other ways." (p.216). Severe problems may exist (see chapter 1). The summary of the results I presented in table 1 does not look particularly promising either. Different methods find very different results. However, if data mining will teach us anything useful, it is worthwhile to pursue in this context.

2.6. Concluding remarks

An important purpose of this chapter was to show the discrepancy between methodological prescription and applied methodology in practice. This discrepancy is strange. If methodology is to have any purpose, it is to make sure the inferences made by applied economists are sound. Prescription and practice should be in line. As this is not the case, either the practice or the methodological prescription is at least partially wrong. Chapter 3,4 and 5 are devoted to examining whether the prescription are right and whether practice is wrong. However, we can also wonder if the fault lies with the goals of the methodologists. As discussed in the chapter, following the strict methodological rules present in the discipline

would result in a neglect of data analysis that may result in very important insights. This seems to be unfruitful from the start. Methodology should be critical where false inferences are made, but should not be careful with making general dogmatic statements about what researchers are not to do. I want to end the chapter with a quote from Deaton (from a different methodological debate) that captures very well where methodology and practice go wrong: [P]ractioners are too talented to be bounded by their own methodological standards" (2010, p. 426).

Chapter 3: Sound Evidence without Theory:
Is double counting really a deadly sin?

"Fitting theory to the facts, probably not such a bad idea after all."

(Howson, 1990, paper title)

3.1. Introduction

There is a very widespread idea about evidence in science that says that a theory can only be said to be supported when the data that supposedly supports the theory was not used in the construction of the theory[11]. If the same data is used in the construction of the theory as well as the evaluation thereof, this is called double counting. In Chapter 1 we saw that an important part of the definition of data mining is that in case of data mining, data is used both in the evaluation and the formation of hypotheses. This is double counting in case both these procedures make use of the *same* data. As this often is the case, like in the three methods discussed in chapter 2, data mining is often an instance of double counting. Double counting is controversial in the philosophy of science in general and the philosophy of statistics in particular. John Worrall (2006; 2010) defends the view that double counting is bad. He has developed a 'use-novelty requirement of evidence' to that effect: if data is not new to the hypothesis, it cannot count as evidence for it. The fact that most data mining is also an instance of double counting is an important reason why it is considered a methodological problem (Spanos, 2000; Mayo, 2008). In this chapter I argue against the view that double counting is problematic. I first provide an overview of the debate on this topic that has recently been revived in a correspondence between Worrall and Deborah Mayo (Worrall, 2006; Mayo, 2008; Worrall, 2010; Mayo, 2010). I review their arguments and contend that Worrall's use novelty requirement is not a convincing necessary condition for evidence. What is important is that the arbitrariness that is associated with ad hoc inference is not necessarily tied to it. There are cases in which ad hoc inference is not arbitrary at all. Moreover, while novelty is surely a virtue of evidence, there are other important virtues too. The virtue of novelty will often come at the cost of the virtue of precision (especially in small samples). I support these claims with some evidence from data mining in the context of the growth economics debate. I argue that double counting is not in itself a reason why data mining is methodologically bad, as methods that double count are sometimes more reliable than its alternatives that avoid it.

3.2. Novelty and prediction

The idea that the data used to construct hypotheses should be different from the data used to evaluate it dates back at least to Francis Bacon (Howson and Urbach, 2005). As in

[11] Steele and Werndl (2013) for instance say how widespread this idea is, and provide many examples of this among climate scientists.

other fields of science, this view is quite widespread in econometrics. In Chapter 1 I discussed the received view of econometric methodology that maintains that data can be useful to test hypotheses but not to develop them: empirical testing and theory development should be kept separate. The *Condemnation Thesis* is based on this view. The view does not only exclude double counting data, but excludes any version of data input in the formation of theories. While this view has been preached, in applied econometrics this is not universally lived by (Magnus, 1999). For instance, the econometrician Jan Magnus writes that there are too many practical problems to make the "top-down" approach fruitful for applied econometrics, and in practice "no applied economist proceeds in this way." (Magnus, 1999, p. 62). In order to construct models, researchers will often simply have to rely on the data. Methodologically speaking, the view has been criticized too. Campus et al. (2005) make the point that it has to be possible for researchers to learn from the data in case they are not sure about the structure of the phenomena they are describing: "one does not need to know all the answers at the start." (p. 1-2).

In the philosophy of science the idea that one can only use data to test purely a priori formed theories is quite out-dated. Consider for instance Hitchcock and Sober: "Nobody thinks that good theories should be constructed a priori, ignoring existing evidence (…). Einstein's General Theory of Relativity, for example, was specifically constructed so that it would closely agree with Newtonian gravitational theory in those domains where the latter was known to match the data well." (2004, p. 6). Ignoring existing evidence in order to make sure that the theory formation and data evaluation are independent seems silly. The quote by which I started this chapter points this out nicely.

A more modest version of the view that theory and evidence should be separated is more popular. Worrall (2006, 2010) accepts that letting theories be inspired by the data is quite fine. He approvingly cites a lecture with the title "ad hoc is not a four letter word". At the same time, he presents a slightly different thesis: theories may be inspired by old evidence, but they cannot be supported by it. New evidence is required for evidential support of theories. While the very strict view has not been defended lately as far as I am aware of, Worrall's account would still rule out most data mining practices.

Recently, Worrall has defended his view against an attempt to refute it by Mayo (1996; 2008). The Use Novelty (UN) requirement, or – as he himself likes to call it –the UN Charter, is neatly summarized by Mayo (2010):

"For data x to support hypothesis H (or for x to be a good test of H), H should not only agree with or "fit" the evidence x, x must itself not have been used in H's construction." (p.156).

Worrall motivates the UN charter by referring to the problem of accommodation, a worry that was itself motivated by the Duhem problem. In case of ad hoc inference, formed hypotheses necessarily accord with the observed data, however, the fact that evidence can be accommodated by a theory does not necessarily speak for the theory. Pierre Duhem noted that a theory cannot be tested by itself, but is always tested in combination with a number of auxiliary assumptions. Hence, whenever a theory does not appear to accord with the data, one can ascribe the failure either to the theory itself, or to the auxiliary assumptions. Consequently, when the data does not accord with the theory, one can always argue that the

true reason for this was not a flaw in the core of the theory, but a flaw in one of the auxiliary assumptions. As long as this is the case, it is possible to accommodate any evidence within a theory by changing some of the auxiliary assumptions. This seems arbitrary. For instance, early Marxists believed that the capitalists would oppress the ever growing lower classes who at some point would take up arms and revolutionize society. When this did not happen, some Marxists maintained that their initial theory was right, but that the only reason no revolution broke loose in capitalist countries was that Marxists made the capitalists aware of the dangers of revolution, which motivated a better treatment of the lower classes. In this case, the general theory (a revolution will occur in capitalist countries) accommodates the evidence (no revolution has occurred yet), by changing an auxiliary assumption (this will only occur if the capitalist do not change their ways). By being liberal with changing the auxiliary hypotheses, it becomes very difficult to find a way in which a core theory can possibly be refuted. The UN charter is intended to avoid this problem. Prediction will have to be based on both the core theory and auxiliary assumptions. The charter requires that the researcher formulates his auxiliary assumptions beforehand, together with his core theory, such that either the whole set is rejected, or the whole set passes a criterion of confirmation. Thereby, living by the UN charter ensures that no auxiliary assumptions are changed to accommodate the evidence within the general theory. In this chapter I ask whether the Duhem problem really is so problematic that all ad hoc inferences are to be considered invalid pieces of evidence.

While many of the opponents of the UN charter side with the idea that prediction is often a very convincing piece of evidence, the controversial part of the charter is that it denies that non-novel evidence can be seen as evidence at all. Steele and Werndl make a useful distinction between absolute confirmation and incremental confirmation. While many may agree that the level of confirmation that predictions can achieve may never be achieved without it (see for instance Hitchcock and Sober, 2004), it remains controversial whether double usage of data cannot lead to the smallest amount of incremental confirmation.

There are alternative views on evidence that maintain that novelty is a virtue of evidence, without requiring that all evidence is novel. Hitchcock and Sober (2004) present two different versions of predictivism: the view that novelty is important, but not necessary, for evidence. Firstly, strong predicitivism states that predicted evidence is always better than unpredicted evidence. Secondly, weak predictivism states that prediction is often a bearer of epistemic qualities, such that predictions are often stronger evidence than non-predicted evidence. These two positions do not exclude the possibility that non-predicted evidence can still provide support for theories or hypotheses, even though they acknowledge the epistemic virtue of predictions[12]. Use novelty may be a virtue of evidence, but not a requirement. Non-predicted evidence may confirm hypotheses in an incremental sense: they may make one justifiably more confident in the truth of a theory or hypothesis.

[12] An interesting note on the advantage of predictions over accommodated evidence is the claim that predicted evidence is not only more reliable, but also gives some confidence in realism of theories. In the methodology of economics this idea is central in Uskali Maki's (2009) interpretation of Friedman's (1953) methodological essay. He argues that Friedman's positive attitude towards predictions rather than truth of assumptions in models can be interpreted as a realistic attitude towards science rather than an instrumentalist one, because prediction test theories "indirectly". Recently, Saatsi and Vickers (2010) argued that it is false to take successful predictions as a sign of reality of theories, by citing Kirchhoff's diffraction theory of light, they show that theories that are false can still sometimes predict better than theories that are true.

3.3. Vetoing the UN charter

While the idea of the UN charter is very appealing, it has been argued that it obscures the real issue: accommodation is not the problem, *arbitrary* accommodation is (Howson and Urbach, 2005; Mayo, 2008, 2010). In some cases, opponents argue, accommodated evidence can support theories, while at the same time some cases of arbitrary evidence may not be accommodated. Hence, use novelty is neither necessary nor sufficient for reliable evidence.

Worrall's UN charter has been criticized from very different perspectives. For instance, Deborah Mayo (2008, 2010) criticizes the requirement from a classical, frequentist, perspective, while Colin Howson (1988; Howson and Urbach, 2005), and Katie Steele and Charlotte Werndl (2013) argue against the UN charter from a Bayesian perspective. All argue that ad hoc inference is very often arbitrary and arbitrariness is bad. However, there are also cases in which ad hocness is not arbitrary, and in these cases, ad hoc inference may be perfectly warranted. In an unpublished talk that Iaonnis Votsis gave in the conference of the Dutch Society for Philosophy of Science the arbitrary cases are aptly called "post hoc monsters". Strong data mining would be an example of such a post hoc monster: arbitrarily fitting hypotheses to the data. In this section I discuss the objections in some detail, and explain why not all ad hoc hypotheses are post hoc monsters.

One objection against the charter is particularly interesting because it appears in the work of both Howson (1988) and Mayo (2008). The objection comes in the form of a counterexample against the charter that is both very simple and convincing. Consider a bowl of red and white marbles of which one does not know the relative frequency. Now someone takes out all the marbles, counts them, and forms the following hypothesis: 50% of the marbles are white, and 50% are red[13]. Now, it seems obvious that this hypothesis is very well supported by the data (maximally supported even). However, as the data was used in the construction of the hypothesis, it seems to go against the UN charter to take it as evidence for the hypothesis.

Worrall (2010) has countered this objection by arguing that this case is not a good counterexample, because it follows analytically that the number of red marbles divided by the total number of marbles results in the relative frequency. This is not what is being tested in the examples. The "experiment" is a simple execution of a method to come up with the value, one already knows analytically to be defined as the mean. The underlying theory in this case is that the ratio of red marbles is the sum of red marbles divided by the total amount of marbles. This is not tested by the observation of the red and white marbles. He refers to this particular instance of confirmation as *deduction from the phenomena* and argues that there is no reason to drop the UN charter because of this. He does slightly rephrase his charter due to this critique by means of the following formulation: "Using empirical data e to construct a specific theory T' within an already accepted general framework T leads to a T' that is indeed (generally maximally) supported by e; but e will not, in such a case, supply any support at all for the underlying general theory T" (2010; p. 143).

[13] Mayo's example is slightly different. She discusses a professor who wants to know the average SAT score of her class. The best way to do this, as she argues, is to ask them all, add the values up and divide it by the number of students.

Deborah Mayo responds by arguing that this is an arbitrary change of the main idea behind the UN charter (2010). As a new counterexample, she uses the case of a murder investigation in which there is a group of suspects, and the DNA of the murderer is available. Linking the DNA to the suspects, is a good method to find evidence for who the murderer may be, even though it does not satisfy the UN requirement. In this case, the conclusion does not follow analytically. Mayo proposes an alternative account of confirmation. According to her, a piece of evidence can confirm a hypothesis if there would be a reasonably high chance that the data would have been different if the hypothesis would be false. She calls this property of a test severity. If a test procedure is indeed severe, a hypothesis that passes this test can be said to be supported by the evidence.

Howson (1988; 1991) has a very different account of evidence. As a Bayesian he argues that the problem of confirmation boils down to how well evidence supports theories, and how strong our prior belief was in the hypothesis. A piece of evidence used in the construction of a hypothesis can indeed support that very hypothesis, but it relies both on the plausibility of the theory and its alternatives to what extent the theory should be accepted. An interesting case from his book with Peter Urbach (2005) is the work done by Adams and le Verrier to incorporate new evidence about the orbit of Uranus in the Newtonion framework. When Uranus was discovered it seemed to go into a very different Orbit than Newton's theory would predict. The discrepancy with the prediction grew every year. Both Adams and Le Verrier developed an account that incorporated the evidence of Uranus in the Newtonian framework by postulating a new planet. While the work was ad hoc, it was very convincing and was indeed confirmed with a prediction a few years later. However, Howson and Urbach argue, there was already plenty of evidence for their theory, before this prediction. In fact, before Neptune was indeed observed, many scientists were announcing a new planet to be discovered in the near future. According to Howson and Urbach, this ad hoc move – to incorporate the newly observed movements of Uranus in the Newtonian framework – was the most plausible of all the alternatives that were available and was therefore not scientifically unwarranted.

A further discussion on the Bayesian side is provided by Steele and Werndl (2013). They approach the problem from a practitioner's point of view. They observe that climate scientists both maintain that double counting is bad, while at the same time many scientific practices in published papers fall under the definition of double counting. One such instance, for example, is calibration: the estimation of parameters in a model that is later evaluated on the same data. For instance, if theory tells us that a certain relation between two variables is quadratic, we still do not know, what the specific parameters are of the relationship. Calibration boils down to maximizing model fit to determine which parameters are most realistic. This is thus a practice of using all the data in the formation of an hypothesis, while the same data is used for the evaluation thereof. At the same time, Steele and Werndl argue that this practice is fine. If one is not sure whether a linear or a quadratic model is more appropriate (all other things being equal), the one that fits with the data best is most likely to be correct. Hence, requiring that all evidence be novel, even in Worrall's adjusted formulation, excludes perfectly warranted procedures from providing evidence.

It may be considered a problem for Worrall that if the UN charter is taken very seriously, evidence can only be found for existing hypotheses, and there is no real space for

discovery. However, this would be wrong. A simple solution to this problem is to separate a data set in two parts: one part is kept for the estimation of a model, or calibration, while a second part of the data is kept for testing. This is Worrall's solution for how one is to learn from the data, in case a good theory does not yet exist. A very important point is made in this regard by Steele and Werndl. While Worrall's solution ensures novelty of evidence, Steele and Werndl argue that this results in much less precise estimations of the parameters than in case all the data is used for calibration. Because, in these cases, less data is used to estimate hypotheses, as part of the data is left out of the estimation procedure to use it as evidence later. In other words, the data is used very inefficiently. This applies in particular to fields of science where data is scarce, such as growth economics and, in the example that Steele and Werndl use: climate sciences.

The UN charter has evoked much response, and while it seems a very intuitive idea, there is something incomplete about it in the least. I now consider some examples from the growth literature to examine the effects of novelty in practice.

3.4. Double counting in the economic growth literature

I now turn to assessing the UN charter in light of the specific case of growth economics I discussed in Chapter 2. This case is interesting because of the little data that is available and the large amount of proposed important variables. The researchers in this field have the choice between not using the data at all, or mining it. The three methods discuss in chapter 2 all double count the evidence. Success of these theories to get it right is therefore an argument against the UN charter. Interestingly, there are good arguments for the claim that, at least under some conditions, data mining does get it right. Moreover, there is at least one data mining method that does not double count the evidence to arrive at their results. By comparing this method with data mining methods that do double count the data, we can get some insights in the isolated effect of double counting on the epistemic value of the evidence in the data mining context.

3.4.1. Can double counted evidence get it right?

From the UN charter perspective, the main solution to the problem of model uncertainty such as the growth economics case is to split up the data into cohorts. Because data is very scarce in the case of growth economics, the cost for the precision of the estimates in case the data is split up is particularly high. The trade-off in is that the more countries are taken out of the first cohort into a second test cohort, the more scarce information is lost. And while the evidence of the novel data points may be convincing, the resulting knowledge is less precise.

None of the three data mining methods discussed in Chapter 2 actually splits the data up into cohorts[14]. By not doing this they breach the UN charter. Leamer's **EBA**, and Sala-i-

[14] Hoover and Perez' (1999) *Gets* algorithm actually takes a subset of the data, but does so with the purpose of detecting structural breaks. The data used for this overlaps with the data used for the estimation, and is thus still used twice.

Martin's **BMA** both use the robustness property of variables as evidence for their importance in explaining growth. In figuring out which variables are in fact robust, all the data is used. The same accounts for *Gets*, in which the final model is constructed by using all the data in the sample. Could these methods possibly result in reliable evidence?

Yes they can. That is, if we believe the arguments from Hoover and Perez (1999; 2004) and Hendry and Krolzig (2003) who have defended the *Gets* approach. One argument is particularly interesting. Hoover and Perez (1999; 2004) test case the *Gets* methodology on a simulated data set. In a first instance (1999), a completely arbitrary data set is used containing a known true model. The search algorithm was used to search the data, and find this model. In general, they find that approximately slightly more than 5% of their found variables are wrongly included in the found model, while very few variables that do not belong indeed end up in it. In a second paper (2004), they use the data set from the growth economists, with 36 variables for 107 countries. They simulate some true models and perform the same tests. Like in their first experiment, they find that the algorithm performs particularly well in identifying the true variables.

In light of the UN charter this finding is odd. If double used data could not result in reliable evidence, then methods based on using the data in this matter should not do so well. This argument may be played down by arguing that the success of *Gets* in the experiment is not due to the method itself, but simply to the set-up of the experiment that made the method perform well. Perhaps outside of the simulation environment this impressive performance breaks down. Even if this is true, then it is still the case that the simulation experiment shows that at least under some conditions, double counted evidence may result in reliable evidence.

This is bad news for supporters of the UN charter: ad hoc inference leads to perfectly good evidence in the context of simulations. Breaching the charter may lead to reliable results. However, it does not yet say anything about the predictionist views on evidence. I discuss another example below that shows that at least the strong predictionist view does not seem to apply to the data mining context either.

3.4.2. How good is prediction really?

In some definitions of data mining, such as that from Stan du Plessis (2009), double counting is taken to be inherent in data mining. While all three methods described so far are both instances of data mining and double counting, there is at least one version of data mining that does not fit this bill: Relevant Transformation of the Inputs Network Approach (henceforth RETINA; White, 1998; Perez-Amaral et al., 2003). In the algorithm of this method, the data examined is indeed split up in a number of different cohorts, namely three. These cohorts have different purposes. The first is used to estimate models that are tested on the second cohort. The models that predict best are then improved by making adjustments that improve the predictability in the third sub-sample. The method is intended to provide the best possible forecasting model. This method also used properties of the data to construct hypotheses and find evidence for them, but in doing so it does not double count any data.

Interestingly for our purposes, a similar simulation methodology to that of Hoover and Perez (1999; 2000) has also been used to compare the RETINA method with the *Gets* method.

Perez-Amaral et al. (2004) and Castle (2005) find that these methods perform remarkably similar in general. However, while RETINA makes better forecasts in large samples, *Gets* performs much better in small samples. Perhaps this is not very surprising. In small samples, the costs of splitting the data set into cohorts for the precision of the estimates are much higher than in large samples. This is important in case of growth economics, because the available data is very limited. Splitting this data up into three cohorts, and using only one to estimate the model parameters on, will come at a high cost of the precision of the estimates. Much of the information in the data will be lost.

In case of growth economics, there are not only very good reasons to assume that double counted data may result in good evidence (Hoover and Perez, 2004), but there is further evidence that it is in fact a better way learn from the data than adhering to the strict UN charter. Recall, the strong predictionist view on double counting is that predicted evidence is always better than underpredicted evidence. This example shows that this is not necessarily true, especially in small samples. The costs of estimation precision may exceed the benefits of avoiding post hoc inference. Prediction may still be a virtue of evidence, but may not always be the most important one. In terms of the Hitchcock and Sober's taxonomy, this is only compatible with weak predictivism: novel evidence may generally be superior to accommodated evidence, but only because novelty ensures that evidence is not *arbitrarily* accommodated. If we know that other methods also do not arbitrarily accommodate the data, it is not necessary that the method that does not double count the data is the preferable one.

To conclude, novelty appears neither to be a necessary condition for good evidence, nor is novel evidence always superior to non-novel evidence. Novelty may be a virtue of evidence, but not necessarily the most important one.

3.5 Summary and some conclusions for data mining

The central idea behind the UN charter is that post hoc inference is arbitrary, and the only way to avoid the arbitrariness is to use predictions as evidence. Worrall in particular is very sceptical about ad hoc inference. However, many have argued that this scepticism is much exaggerated. Ad hoc does not always mean arbitrary. Examples, such as the discovery of Saturn, DNA tracing of criminals and calibration, seem to point out that the UN charter is too strict.

In an attempt to examine the problem in the context of data mining, I looked at evidence on double counting in simulation experiments. Some of these results were promising. In fact, it turns out that the data mining methods that use all the data for both estimation and formation as well as evaluation do better in data samples that are like the cross-country growth data than methods that separate the data for these purposes. This goes against the UN charter, and even against strong predictivism.

What does this mean for data mining? The lesson is simple. The claim that data mining is double counting is one that is taken to be an important objection against data mining. The argument made in this chapter shows that double counting is not something to be particularly worried about in general, and especially not in the case of data mining. Accommodation in general is not something to be worried about, but *arbitrary*

accommodation is. Hence, strong data mining should be considered to be bad practice. But this is because strong data miners arbitrarily accommodate evidence, and not because it double counts the data in general.

It may be objected that my reasoning in the second part of this essay has been somewhat circular. I have used arguments from the data mining context, to argue against the UN charter, which is an objection against data mining. However, the example where RETINA and *Gets* are compared allows us to isolate the double counting aspect in the data mining context. This aspect itself did do no harm, in fact, it did much good in the case of small samples. I thereby have shown that the objection that data mining is double counting is not a very good objection against data mining. Nevertheless, for an example against the advantages of novel evidence in another context, I present an example as a post-script.

The upshot of this chapter is well supported: non-novel evidence may be very reliable evidence. However, in applied research, a message like this may not be very useful. As we saw in this chapter, different virtues of evidence may have to be traded off. In our case, precision versus novelty. What is really needed to guide methodological prescriptions for applied researchers is to know how important evidential virtues like these are relative to each other. This work remains to be done.

Interlude to chapter 3:
Do Sagittarians break their Arms more often?

Reconsider for a moment the astrology example from chapter 1 (Austin et al., 2006). The researchers studied the 10,000,000 inhabitants of Ontario and started to test a large number of relationships between medical problems and astrological signs. They found 24 statistically significant relationships. The researchers started without any specific hypotheses, ended up with 24 that were constructed because they fitted the data well, and the p-values were used as (admittedly false) evidence. Hence, this is double counting. Recall that they divided their data set up into two cohorts, and used one part of the data to find statistical relationships, and one part to test them on. Note that there exist a strict separation between formation and testing of hypotheses, and hence it is not a breach of the UN charter. Oddly enough, out of the 24 hypotheses, 2 still remained significant in the second cohort. The prediction that Sagittarians seemed to break their arms more often, based on the first cohort, worked well in the second cohort too. The researchers conclude that predictions do not provide any promise that spurious results will not be found.

A danger of focussing on the UN charter is that in light of the UN charter novelty is seen as the most important quality of evidence. In this case this would be grossly wrong. The authors argue that a better way to learn about the relationships between astrological signs and disease patterns is to test a large set of hypotheses without splitting up the data set and to correctly take account of the large set of tests that were run. When the data analysis is conducted in this way, the dubious astrological sign-disease patterns disappear. Like in the examples in the context of growth economics, we see that a better way to learn about the data is to carefully exhaust all the information from the data, rather than to split it up and use one part to form hypotheses and another to confirm them.

Chapter 4: The Original Sin:
On quantifying search in economics

"[D]ata mining is not a sin. After all, how can we avoid it? It is a sin, however, to ignore the effects of data mining."

Jan Magnus (2002, p. 607 [original emphasis])

4.1. Introduction

In a talk given at the INEM conference in Rotterdam in June 2013, Deirdre McCloskey discussed how she used to do econometrics when she was studying economics. In order to estimate a regression equation, much work had to be done in order to find the data, put it in the right form, put it into a punch card, and wait a couple of days to get the results. Since then, technology has developed. And presumably this is good news for the economics profession. Today, I can both find data sets and use them to estimate complex regression equations on in a matter of minutes, if I know where to look. And thus, the cost of doing statistical analysis has dropped significantly.

Unfortunately, this fact does not only have positive consequences. It also complicates matters. The method of statistical inference that is used most in econometrics is not developed for inference in the context of many statistical tests, but was intended to teach us what we can learn from single experiments. This is emphasized, for instance, by Jeffrey Wooldridge, the author of my undergraduate textbook: "The results (…) we derived for hypothesis testing, assume that we observe a sample following the population model and we estimate that model *once*." [15] (p.678, my emphasis; a similar statement is found in Hollanders, 2011). This brings along an odd problem. The fact that it becomes easier to do statistical analysis makes it more likely that a statistical finding is a result of pretest bias. And thus, the interpretation of such analysis has become more complex. Unfortunately, this has not been recognized by the economics profession, that still uses the same statistical analysis that was used in times when it did take much time and effort to estimate statistical models. In this chapter, I discuss this problem.

It is often argued that data mining is bad because it distorts the interpretation of statistics due to multiple testing: running more than one test, and using the results of some of these tests as evidence (discussed in chapter 1; in particular Lovell, 1983). Multiple testing brings along the danger of pretest bias. The *Condemnation Thesis* (**CT**) is based, for a large part, on worries about the statistical interpretation of results that come out of a search procedure. An econometrics textbook I used in my courses illustrates the point nicely. It warns its readers for running multiple test to arrive at the best specification as "[t]he probability of making incorrect choices is high, and it is not unlikely that your 'model' captures some peculiarities in the data that have no real meaning outside the sample" (Verbeek, 2008, p. 59).

[15] The results he refers to are the F- and t-distributions for model fit and deviation of a parameter from the null hypothesis respectively.

At the same time, I take it as part of the received view that it is not the search by itself that is problematic, but that it is only problematic when researchers do not report all their results that were used for the search. Mayer (2000) or Hollanders (2011), for instance, take the view that data mining is only considered a problem because researchers report their results selectively. The view is worded nicely by Leamer: "Sinners are not expected to avoid sinning, they need only confess their errors openly" (1978: preface).

However, in this chapter I show that the thesis that all relevant results need to be reported in order to correct for pretest bias is false, because there is no way to identify what all relevant results are. In the former chapter I argued that double counting, which is often regarded as a serious problem for data mined results, is in fact not by itself a problem at all. In this chapter I take another strategy. I argue that the problem of pretest bias, which is often taken to be a major problem for data mining, is not exclusive to a small set of cases which are deemed bad data mining, but is omnipresent in all statistical testing in economics. Any statistical result in economics is to some extent the result of multiple testing. And therefore, the classical interpretation of statistics is distorted in case of statistical analysis in economics. Consequently, it is not very meaningful to condemn data mining for this reason. The argument in this chapter is mostly negative: the objection that data mining distorts the interpretation of statistics is not very useful to distinguish warranted and unwarranted cases of inference, as almost all statistical inference in economics will be subject to this objection. In the next chapter I take a more positive outlook and explore alternative ways to the question how we can learn from statistics if we know it is a result of search.

At this point, I can give away one consequence of the argument. Data mining is often considered a moral flaw of applied econometricians. Searching through the data to find good fitting models is considered bad practice. At the same time, it is common practice in the econometric profession. If the argument in this chapter is correct, we can see that the key problem of data mining is not the disability, or unwillingness, of applied econometricians to report all their results, but rather a deep methodological one. If we see data mining as a sin, it is like the original sin: it is bad, but we cannot avoid it. It makes all econometricians bad from their birth in econometrics 101 on. But perhaps it is better to see data mining otherwise, and take serious the words of Edward Leamer: "unavoidable sins are not sins at all." (1978, p. vi)

4.2. Search, Selection and Pretest Bias

The problem of multiple testing is best explained in (though it is not limited to) the classical statistics framework. The classical framework is greatly reliant on p-values: the classical measure of the conditional probability that the evidence, E, was brought about, were hypothesis H true (i.e. P(E|H))[16]. If this value is particularly low (smaller than 5%), then it is considered exceptionally rare, and taken as evidence against the hypothesis. A small value of P(E|H) can be taken as evidence against that very hypothesis, H[17]. It is this proposition that is problematic in the context of data mining. Because, P(E|H) can only be reliably estimated

[16] In reality, classical p-values measure something slightly different (Keuzenkamp and Magnus, 1995): the chance that something more deviant from the hypothesis than the observed came about, were they hypothesis correct.

[17] This is the case in both the classical and the Bayesian framework.

when there is no search effect: finding spurious statistical relationships in the data due to running many tests. In case of data mining, evidence is selected exactly for its evidential value, be that a low p-value, or another indication of a good fit with the data. And in such cases of selection or search, statistics such as p-values do not truthfully represent the strength of the evidence anymore.

This can be illustrated by means of a toy example. Consider a person who is interested in testing hypotheses in the following form:

H_i: person i does not have a special ability to throw double sixes when he throws a pair of dice.

If this hypothesis it is true, and let us suppose it is true for whoever person i is, person i will be unlikely to throw double sixes. However, the researcher now goes looking for evidence. He asks a large number of people to roll a pair of dice. At some point he will be expected to find someone – let's say, person 36 – who throws double sixes on a first throw. The evidence that a person throws a double pair of sixes, if H_{36} is true, is quite low (2.8%), and hence, he may conclude that the H_{36} is likely wrong and that person 36 has a special ability to throw double sixes. However, this is obviously false. In fact, the chance of finding evidence against one of the hypotheses is not that low. It is in fact expected. However, it does seem particularly low if the rest of the results are not taken into consideration. Finding the result for the specific hypothesis (H_{36}) was unlikely, but it was likely that an unspecified piece of evidence against one of the hypotheses (H_i) tested was found. Not taking this into consideration, and concluding that it is plausible that the 36th person indeed has a special ability to throw double sixes on the basis of this experiment is wrong. We can call it the *fallacy of neglecting search*.

This is a whimsical example. However, the same logic applies to hypothesis testing more generally. Hypothesis testing is based on the law of the large numbers (LLN). The LLN says that for any probability distribution[18], such as one describing a statistical relationship between two variables, we expect that if we take a random sample, the mean of the sample is normally distributed. From this law we learn that if a set of samples becomes large, we do not only expect to observe values closely around the mean, but also deviations from the mean. Consequently, if 20 true hypothesis tests are conducted, it is expected that one rejection is nevertheless found at the usual significance level of 5%. While this is not a very realistic case, the same mechanism at play in this example affects the found results in more realistic cases. In chapter 1 I mentioned Friedman who worried about this effect in the work of Tinbergen. If a researcher selects variables in the model because they correlated well with the dependent variable, the model is likely to fit very well too, but one should be highly sceptical about the meaning of the results. The statistical tests are likely to look good, but the quality of the statistical tests will represent the perseverance of the researcher, rather than epistemic reliability of the results.

P-values and hypothesis testing are central to statistical inference in the classical sense (see chapter 1). Classical inference starts from the assumption that a low value of the

[18] With the exception of slowly converging ones.

probability that the evidence was brought about under a hypothesis, P(E|H), is a sign of falsity of the hypothesis, H. We can express it in the following reasoning scheme.

Low P-values → low P(E|H) → Exceptional instance, given H→ implausibility of H

We can call this the inferential interpretation of p-values. In the examples, this interpretation breaks down. In order to avoid the fallacy of neglecting search, we need to be able to avoid pretest bias in cases of search.

The reason this problem is called the problem of pretest bias is because it refers to testing a statistical relationship, which has already been selected for fitting the data well. Spanos (2000) provides a nice illustration of this bias. According to Spanos, running tests to see which one work well, statistically, is like shooting at a wall, drawing a bull's eye around the bullet whole, and calling oneself a sharp shooter. Only a sharp shooter could hit the center of the bull's eye if it was drawn in advance, but any person could get a shot in the center of the bull's eye if it was drawn on post hoc.

Lovell (1983) shows that this effect is very real: if classical methods are used multiple times, many false inferences will be made. As Hoover (1995) points out, there exist many accidental correlations in the world, and it is likely that there are a large number of macroeconomic variables that correlate with important economic series, such as the United States' GDP, simply by chance. If we search for correlations in the data, we are likely to find many that are simply due to chance. In fact, data mining is likely to produce results that perform better according to the statistical criteria, such as model fit, than models that describe the true structure. This is due to *overfitting*. The true model, even if it is known from the start, never performs perfectly due to random error. If one has enough data, it is always possible to find a model that fits better than the true model, even though it does not describe the true structure. In short, data mining makes it very likely that a model is found that performs very well statistically, but provides little presumption about its truth.

The problem is now phrased in terms of the classical, frequentist, framework. One may wonder if the problem is not simply solved by moving to the main alternative framework: the Bayesian one. Unfortunately, this is not the case. The difficulty that multiplicity of tests poses is that in case of search an exceptional value does not necessarily represents a low probability of observing the evidence were the hypothesis correct, P(E|H). This probability also plays a crucial role in Bayesian reasoning. Bayesian reasoning is based on Bayes' theorem: $P(H|E) = \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|H_a)P(H_a)}$, where H_a is the alternative to H. The probability P(E|H) plays an important role in this function. And, independent of being a bayesian or frequentist, if one searches long enough, low values of P(E|H) will be found, and the posterior probability, P(H|E), resulting from Bayes' theorem, will misrepresent the appropriate confidence one should have in the hypothesis because P(E|H) is misrepresented[19]. Bayes' theorem may provide fruitful points of departure for solutions to this problem. For

[19] It may be argued that in the Bayesian case, evidence is always updated, and hence, searching for evidence is never unwarranted. This is a controversial point of discussion in the debate on stopping rules (cf. Mayo and Kruse, 2001; Steele, 2012). I think this position is wrong, especially in case of data mining, because updating only occurs if data is gathered on the same hypotheses, while data mining may occur over different hypotheses, as I discuss later in this chapter in more detail.

instance, both Leamer's approach as well as the **BMA** apply Bayesian reasoning in order to deal with this issue. The efficacy of their solutions will be discussed in chapter 5. What is important as this point: the problem of search effects is not solved by moving to a Bayesian framework.

4.3. The Corrective Condemnation Thesis and the Inevitability Thesis

As the discussion above shows, a great deal of the problem of multiple testing is that other tests are not considered, while they are relevant in the assessment of the evidence. If the experimenter in the dice roll example had taken into account the fact that the one significant finding was simply one test out of 36, he would not have considered it as evidence. Thomas Mayer defends the view that unwarranted data mining is mostly an informational problem (1980; 1992; 2000). If all researchers would report all their findings, then the search procedure would be transparent and there would be no methodological problem related to data mining. I take the most common view on this topic to be very similar. Search by itself is not a problem, but failing to report it is. Kennedy (2002) for instance, writes – in an advice for applied econometricians – that data mining is not by itself bad, but "be aware of the costs" (p. 577). This is the same attitude that Leamer takes to be accepted view among economists (1978): "Sinners are not expected to avoid sins, they only need to confess their sins openly" (p. vi). Call this version of **CT** the *Corrective Condemnation Thesis* (henceforth **CCT**): data mining is bad practice only in so far it done without reporting it.

There is some debate about this matter though. Hoover (1995; and Hoover as quoted by Pagan and Veall, 2000) is particularly sceptical of the view that all results should be reported in order to avoid problems with respect to data mining. Because, he wonders, what would one do with all the reported results? Pagan and Veall (2000) agree that there may be "too much of a good thing" (p. 213) if it comes to reporting results. In the next section I want to examine the possible ways in which this problem can be dealt with. I first want to examine some sceptical views about **CCT**. Without going into the details of specific correction mechanisms, which the following section deals with, I review some authors who have taken very sceptical positions about this view (in particular Greene, 2000; and Hoover and Perez, 2000). They argue that either correction methods cannot, in principle, properly quantify the effect of multiple testing, or they cannot do so for practical reasons (or a combination of these two). We can call this view the *Inevitability Thesis* (henceforth **IT**) as it maintains that the pretest bias due to multiple testing is not something that can be avoided by correcting properly for them.

An important note on terminology is that in these debates data mining is always put on par with pretest bias. Recall that in the introduction the thesis was discussed that data mining is unavoidable (Leamer, 1978; Frank Denton, 1985; Halbert White, 2000). By this they mean that pretest bias due to multiple testing is unavoidable. For reasons of conceptual clarity I have formulated **IT** in terms of pretest bias rather than data mining. However, saying that unintentional data mining is unavoidable has a similar meaning to saying that pretest bias is unavoidable.

Hoover and Perez (2000) and Greene (2000) argue that even if the rules set out by the early econometricians are followed, it may nevertheless be the case that search effects occur due to unintentional data mining. In such cases, the classical interpretation of hypothesis testing is still distorted due to the same kind of bias that is present in case of intentional data mining. Greene worries in particular that search may not be limited to individual researchers working on the same question, but may includee research fields at large (this is also found in Denton, 1985; and Caudill, 1988; 1990; Lo and MacKinlay, 1990; and White, 2000). He puts the problem as follows: "The 'data-mining' engaged in by one researcher is only a marginal contribution to the collective search process. This search process corrupts and invalidates all formal statistical 'tests' conducted with data extending backwards into the past." (p. 221). This pertains in particular to time series, such as the United States' GDP, that are used again and again for the testing of hypotheses. As soon as the results of these empirical procedures lead to new ideas about a certain field of interest, the initial tests become an important factor in determining future research. Hence, statistical properties of the data are used both for the formation of theories as well as the testing of them. According to Greene the history of economic hypotheses needs to be considered in the evaluation of its statistical evidence in order to be fully able to quantify the effects of multiple testing. At the same time though, he asserts that this leads to ridiculous conclusions. For instance, econometrics courses would have to avoid real data in danger of creating a large pretest bias, because the statistical relations found in the course may bias the researchers' ideas on the matter. According to Greene, solid classical inference requires that the hypotheses are only based on "a theory which is constructed without prior knowledge of the world." (p. 221). Obviously, this is impracticable in non-experimental fields of economic research. It is both necessary and straightforward that economists who develop theories do take the evidence into account that is available at the point of development. Consequently, Greene takes the position that multiple testing may be corrected for in principle, but not in practice. He therefore argues that the classical rules of statistical inference only apply to experimental findings and empirical findings that use out-of-sample data to test their results on[20].

Hoover and Perez (2000) have similar worries: To avoid a pretest bias in classical testing "one must establish a measure of the amount of search and keep track of it. Yet, typically economists do not know how much search produced any particular specification, nor is the universe of potential regressors well defined. We do not start with a blank slate." (p. 199). Like Greene, Hoover and Perez argue that in order to make avoid pretest bias, it is necessary to keep track of all the relevant research that has led to the specification that a researcher ends up with. They continue: "A specification such as the Goldfeld money demand equation has involved literally incalculable amounts of search. Where would we begin to assign epistemically relevant numbers to such a specification?" (p. 199). A particular problem with this is that samples often overlap. Data in the form of a pattern that may have inspired a hypothesis 20 years ago will, in addition to newer evidence, remain to be used to test the hypothesis. Hoover and Perez find that separating theory formation and statistical testing is not a realistic possibility. Moreover, keeping track of all the data influences that inspired a

[20] Out-of-sample testing means that the data set is split up in two parts (see Chapter 2 and 4). One of these parts is used to estimate a statistical model, while a second part is used to test the findings of the research on the first part.

hypothesis is similarly impossible. Like Greene, Hoover and Perez assert that **CCT** leads to untenable conclusions. According to Hoover (1995), the problem is not merely practical. He voices doubt about what is to be done with the information, even if it is reported.

In short, while the **CCT** view is widespread, it appears to have problems. A number of authors point out that all the relevant information needed to correct for multiple testing is much more than that of the single researcher, and that there would be large practical difficulties in keeping track of all the search conducted in order to arrive at a single interesting test. If we accept their argument, we must conclude that the crucial problem caused by multiple testing, pretest bias, cannot be avoided by sticking to the rules.

4.4. Correction methods in classical statistics

While the insights by Greene and Hoover and Perez are fruitful starting points for thinking about this problem, their discussions on the unavoidability of pretest bias are not very detailed. Neither paper spent more than a paragraph on the topic. In order to examine **IT** I consider what kind of information would indeed be necessary to avoid pretest bias. Outside of economics, much has in fact been written on the topic of multiple testing (e.g. Hochberg and Tamhane, 1987; Benjamini and Hochberg, 1995; Abdi, 2007). In classical statistics there are a number of orthodox solutions to this problem. In economics it is sometimes mentioned that in order to solve data mining, one needs to apply these methods (e.g. Hollanders, 2011). However, I am surprised to find that the issue has not been discussed much. In this chapter take up this issue. The analysis of the requirements for classical correction methods will in fact lead to a rigorous defence of **IT**. While some of the conclusions that I draw have been drawn by statisticians in other debates (in particular Hurlbert and Lomberdi, 2012), they have not been discussed in the context of data mining.

The two main approaches tackle the problem of multiple testing slightly differently. Both the Family Wise Error Rate (FWER) and the False Discovery Rate (FDR) are ways to quantify the effect of multiple testing on inference. Firstly, the FWER is a way to quantify the probability that at least one of the tests in a set of tests falsely rejects and to fix this probability at the conventional α. Specific methods to calculate this probability are the Bonferonni and Šidák corrections (Abdi, 2007). They are based on the following intuition. If the chance of making one type I error per test, that is wrongly rejecting a true hypothesis, is 5%, then the chance of making one type I error in 10 tests is much higher (about 40%). In order to fix the value of making at most 1 type I error in the set of tests, one should maintain stricter criteria in all the singular cases. This set of tests is called a family. If α(PT) is the chance of making a type I error in one specific test, and α(PF) is the chance that one Type I error has occurred in the family of tests, and C is the number of tests that belong in the family of tests, than their relation is as follows:

$$\alpha(PF) = 1 - (1 - \alpha(PT))^C \quad (3)$$

For instance, in case an α(PT) is taken of .05, as is standard, and 10 different test are run, the chance that there is at least 1 type I error among them is:

$$1 - (1 - .05)^{10} = 1 - (.95)^{10} = .401 \approx 40\% \quad (4)$$

In other words, the chance of making at least 1 type I error in a set of 10 tests of which all the null hypotheses are in fact true is approximately 40%. The correction is then, that rather than the desired level of significance, the individual tests should use a stricter criterion, namely, the following[21]:

$$\alpha(PT) = 1 - (1 - \alpha(PF))^{1/C} \quad (5)$$

where the α(PF) is set at the desired level. In our example of 10 tests and a desired level of significance for the whole group of tests of 5%, we can calculate the desired stringency of the single tests as follows:

$$1 - (1 - .05)^{1/10} = 1 - (.95)^{1/10} = .0051 \quad (6)$$

It turns out that in order to maintain an overall chance of making a type I error of 5%, we have to test all the individual hypothesis with an α of .51%[22].

The FDR does something different (Hochberg and Benjamin, 1995). Hochberg and Benjamin argue that the FWER is too strict a rule for statistical inference, as it is intended to ensure a small chance of at least falsely rejected hypothesis error, but the risk of not rejecting a false hypothesis, making a type II error, is not considered. They therefore propose to not to focus on how to control the probability of making at least 1 error, but rather on estimating the amount of rejected tests that are expected to be false in a set of multiple tests. For instance, consider a researcher who has run 100 tests. Forty of them are significant under the conventional level (5%). In other words, he has made 40 discoveries, while trying 50 specifications (100 is the family in this case). However, if all hypotheses were true, we would expect 5 of them to pop up by chance. However, in total 40 did pop up significant. Hence, of all our rejected hypotheses, we expected $\frac{5}{40} = \frac{1}{8}$ to be false due to the search effect. The other 35 discoveries are still expected to be correct. The False Discovery Rate is a ratio of the expected number of false discoveries divided by the total number of significant findings. An insight FDR provides is that the problem of multiple testing is more worrying in case few tests are found to be significant (cf. Hendry, 2002). For illustrative purposes I have often assumed in my toy examples that all null hypotheses are true. In reality this is often not the case (see de Long and Lang, 1992). FDR provides an important insight in the nature of multiple testing: the more tests actually are significant, the smaller the chance that they are significant due to chance.

To illustrate the difference with the FWER, let's consider a case in which the 40 tests were all significant, but only marginally. If the FWER would be used to make the testing rules

[21] An important note is that these corrections are appropriate if the test are independent. If the tests are in fact dependent, then the FWER provides a conservative approximation of the appropriate α to be used per test.

[22] This is the Šidák correction of multiple testing. The Bonferonni method is a simple approximation of this: . As you can see, in our example, this would result in a slightly lower criterion (.005)

stricter such that we would be reasonably sure not to make any false discoveries, all the 40 findings would have turned insignificant. Hochberg and Benjamin therefore argue that the FDR is more reasonable. What the example shows is that Hoover (1995; Hoover and Perez, 2000; Pagan and Veall, 2000) asks a sensible question when he wonders what the use is of reporting all results. It is ambiguous what we can conclude from reported tests. However, the ambiguity between these two methods is not the problem I focus on. In the next section I analyse a problem related to a communality of the classical correction methods: they both rely heavily on the concept of a family of tests.

4.5. Families of tests

Consider a researcher who feels that he needs to run multiple specifications in order to find the best one, because he does not have a proper theory to establish this for him. This, for instance, could be a growth economist from chapter 2. He worries about the effect of pretest bias due to multiple testing. He wants to correct for this in order to make sure that this bias does not invalidate his found results. In order to do so, he uses the FWER, but the following discussion would also apply if he would prefer the usage of the FDR. It becomes a crucial part of the correction procedure to keep track of all the relevant hypotheses tested. After all, for both procedures (FDR and FWER) it is crucial to define a family of tests to which one test belongs. The first problem then becomes: Do all the tests run that related to the research that was conducted need to be reported as part of this family? Do, for instance, diagnostics tests need to be put in here too? Or, another question is whether only tests that were run on one data set should be included, or, if two data sets were used in one study, should tests run on both data sets be corrected for? The crucial question is to what extent a reasonable answer exists to the question What is a family of tests? I argue that there is no straightforward answer to this question. There might be a straightforward answer in case of experimental research, but in economics, things are too complex. It is hard to quantify the amount of search conducted to end up with one result. We can call this problem pretest bias uncertainty.

In order to clarify this problem I focus on the concept of a family of tests. The reason for this is that it is the way in which the two most orthodox methods to correct for search conceptualize search effects. The discussion will highlight, though, what kind of reasoning is necessary for thinking about the concept of search, and statistical inference in light of it.

A few constructive suggestions exist in the literature that may help to sort out these questions. Firstly, we can ask ourselves what we mean by the term, or what the term is intended to describe. Tamhane and Hochberg (1987) provide a conceptually clear answer. They define a family as “any collection of inferences for which it is meaningful to take into account some combined measure of error” (p. 7). This is the definition used here. In particular, we can add to this definition that it is meaningful to take into account some combined measure of error, if not doing so may result in pretest bias. However, this definition does not provide any practical answer to questions mentioned above. Tamhane and Hochberg do not want to go into more detail, because they believe that there are significant difficulties with this concept which they feel only have good solutions in particular contexts.

Miller (1981) is one of the few accounts that provides a more substantial, practical, answer to the question what a family is. He argues that while it is not easy to say which tests belong in which family, "[t]he natural family for the author in the majority of instances is the individual experiment of a single researcher" (p. 34). This is perhaps a somewhat simplistic account of a family. It seems somewhat arbitrary to restrict correction methods to a single researcher, or in particular to a single experiment. For instance, it has been heavily criticized by Hurlbert and Lomberdi (2012). They refer to Saville (1990) to make their point: "An experiment is no more a natural unit than a project consisting of several experiments or a research program consisting of several projects."

On the other hand, there is something to this definition. Ronald Fisher developed his statistical method to study experiments (1935), and search effects are quite unlikely to pop up in case of experiments. After all, experiments take time and cost money. It is therefore not easy to search among them. One may easily search in the data from one experiments, but one cannot easily search among different experiments to find satisfactory statistical relationships. There is perhaps something arbitrary about describing families of tests as those that belong to the same experiment, but, Miller is right to say that they are to some extent "natural".

However, in case of (non-experimental) economics, the meaning of a family is much less straightforward. The kind of economic research that I focus on in this thesis, and that is most typical in economics, relies on observational data rather than experiments. If Miller is right to assert that a very natural family is a single experiment for a single researcher, it would still be unclear what a similarly natural family would be in these cases. A natural counter-part of an experiment in observational data analysis is perhaps a regression. Regressions are estimated on the data in order to find an answer to the question how well a certain statistical relationship fits with the data given some controls. Conceptually, this comes quite close to what experiments are intended to do. However, if it comes to search effects there is a crucial difference between experiments and regressions. As we saw in Chapter 2, it is not difficult to run many regressions on a data set (see Sala-i-Martin, 1997 and Ley and Steel, 1999). It is no coincidence that we have never seen a paper with the title *I Just Conducted Two Million Experiment*s. This would generally take enormous amounts of time and money. It is much easier to search among regressions than among experiments. Hence, if experiments are indeed a natural criterion for families of tests, then regressions are not. If a researcher indeed runs 100 regressions and investigates a number of statistical tests per regression, should he, for proper inference, correct for all the other searches if he presents his significant findings or should he consider the regressions separately? The analysis would be greatly lacking if the researcher would focus only on his best fitted regression, without considering the others.

There are other concepts that could serve as a natural interpretation for families in observational research. Mayer (2000), for instance, does not say that a researcher should report all the tests run in one regression, but suggests that all regressions conducted in the construction of a single paper be reported. This is probably the most common view about correction. Jan Magnus (2002) suggests that the researcher reports a logbook of his work. This would make science very transparent if followed by everyone. The scope of the test conducted in one research project is larger than the tests conducted in a single regression. However, as an interpretation of a statistical family it is still not large enough. This is because selection effects happen at the paper level too. The most famous example of this is the

publication bias, or file drawer problem (Rosenthal, 1979; but already discussed as a problem in Sterling, 1959). Consider an economist who, after much statistical search, finally found a statically significant finding of his interest. In his paper he reports all the preceding search. This would be very rigorous. However, it may be the case that a second researcher conducts a similar project, but is unable to find anything significant. They might both write papers about it (though the latter probably would not bother) and only the former would get published. All his individual results may still not be a sufficient collection of tests relevant to evaluate his hypothesis test of interest, because it does not contain all search conducted by his colleague who did not find anything. However, such independent search procedures may still bias the results that roll out of it.

It may sound somewhat strange that research conducted by other researchers affects the evidence found by a different researcher. However, Jan Magnus (1999) notes the very same problem as something he ran into as an applied econometrician. He describes that he, at some point in his career, was working on a paper on capital, labour and energy in the Dutch manufacturing sector (1979). He wonders: "There are two closely related studies, one with Canadian data and one with US data. What do we do with these other studies? Clearly they contain some relevant information, but how can this information be properly incorporated in the Dutch study?"(p.64).

The original file drawer problem was formulated as a problem for meta-analysis (Rosenthal, 1979). Meta-analysis is a method intended to summarize, and aggregate, the results about a certain topic of interest that has been studied in many instances. However, as only significant findings have a good chance of being published one will get a biased picture of the conducted research (e.g. Scargle, 2000; Stegenga, 2011). Interestingly, Stanley and Doucouliagos (2010) recently published a paper in which they examined some popular topics of research in economics (such as the effect of minimum wages on unemployment) for signs of publication bias. As it turns out, the reported results were not distributed as one would expect according to the Law of the Large Numbers in a number of research fields. Stanley and Doucouliagos argued that this was a sign of publication bias.

To provide an example from econometrics, consider a question in financial economics: do stock returns follow a random walk? (e.g. Lo and MacKinley, 1999) This is a tricky question to answer, because it is not easy to say if a certain data series follows a random walk or not. Especially in small samples, random walks may look like trending time series. The efficient market hypothesis, which is the more general version of the random walk hypothesis has been an accepted theory, as it follows directly from assuming rationality of stock traders. If stock traders are all rational, and they are roughly neutral to risk, one would expect all stocks to be priced at such a rate that the expected returns on all of them are similar. Hence, all innovations in stock prices will be due to surprise information, and will be unexpected. Hence, we can expect that the prices of stocks will follow random walks. In the 1990's people started to test this assumption by testing whether there are indded no strategies that would result in structurally higher returns (e.g. Jegadeesh and Titman, 1993; Lo and MacKinlay, 1999). For instance, would buying winners and selling losers outperform buying and selling at random? Many such patterns in the data were indeed found. Interesting for our purpose is the question to what extent these patterns were genuine or due to the search effects caused by so many different researchers looking for patterns. Lo and MacKinley (1990) warn for this

effect. Sullivan et al. (2001) and Neuhierl and Schlusche (2009) go further than this. A common strategy to finding anomalies to the random walk hypothesis was to look whether certain stock returns were different depending on the days of the week or months of the year. Sullivan et al. (2001) assess all these possible calendar rules that one can make and assess whether the reported statistical significant findings are to be expected even if no calendar effects exist. They argue that this is so. Neuhierl and Schlusche (2009) consider a much larger set of complex trading rules that have been examined over the years. They argue that all reported trading strategies lose their statistical significance if seen in light of the overall search. Both papers conclude that reported trading strategies are much more likely the result of search effects than of genuine patterns of profit-rewarding strategies to be found on the trading floors.

In short, the publication bias is real. If one interprets the family of tests to which findings in observational data research belong to be the research done for papers, one misses this bias, and may get an incorrect picture of the amount of research conducted in order to arrive at the finding.

This point is related to a somewhat arbitrary part of Miller's definition of families: the tests conducted by "a single researcher". In many cases, search occurs with more than one researcher involved. For instance, this may happen when a group of researchers investigates the same problem. In research in the field of growth economics, a researcher may investigate the effect of religion on growth, find that there is nothing there, and inform his colleagues about this failed endeavour. His colleagues might therefore be demotivated to investigate this area further. Hence, someone looking only at the search conducted by a single researcher, might overlook search conducted by others that he takes into account.

Lastly, one might suggest that fields of research are good demarcation on what should be corrected for and what should not be corrected for. If a researcher may do work in labour economics and growth economics, then he need not consider searches in the one field if he works in another field. Right? In a one page paper, Gordon Tullock (1959) argues against this idea. Search procedures may happen over different fields of economics as well as different regressions in the same field he argues. While I do not want to defend this claim myself, it shows how hard it is to fully quantify the amount of search behind a single empirical result.

A final note on classical correction for multiple testing in economics is this: so far we have focused on the concept of a family, and its difficult application in economics. However, a further point relates Hoover and Perez' (2000) and Greene's (2000) worries to classical correction methods. Classical correction methods generally assume independence (Abdi, 2007). However, in Chapter 2 we looked at tests conducted about parameters in regression models that all had the same dependent variable (economic growth). Because the same samples get examined a lot in economics, independence is hardly ever a realistic assumption. This point does not relate directly to the concept of a family, it is simply a so far undiscussed further problem with quantifying search and accounting for search effects in empirical research in economics.

Consequently, demarcating search families is conceptually difficult, and perhaps even impossible. We have to conclude that almost any statistical result in economics can be affected by pretest bias. Pretest bias is thus not something that can help us distinguish

between cases of proper hypothesis testing and improper data mining. The problem appears to be quite fundamental.

4.6. Is it really that bad?

So far I have discussed problems. Quantifying the effects of search is difficult. Is there really nothing that could help us quantify search behind statistical findings? The discussion above has almost exclusively focused on the conceptual problems with families of tests. I want to point out one possible solution to the problem. Many take the pretest bias problem to be a plea for theoretical guidance of empirical inference (e.g. Austin et al., 2006). Perhaps theoretical foundations for hypotheses may help to differentiate between expected random errors and plausibly genuine deviations from a hypothesis. Consider our dice example from the beginning of the chapter. A statistician learns about a person who has not only thrown one pair of sixes on a first roll, but a second pair of sixes on a subsequent roll. He is baffled, and takes it as evidence against the hypothesis that he is dealing with a normal dice thrower. Then he learns that the same experimenter also asked 800 other participants to roll the dice. As the statistician knows that 1 in 1296 consecutive double dice rolls are expected to result in consecutive double sixes, he changes his attitude towards the considered hypothesis, and concludes that the consecutive double sixes roller must be a random event. But again, he learns a new fact. This time he learns that the person who was throwing the dice was not a regular participant, but was a magician. Moreover, he brought his own dice. The statistician now concludes that he is not dealing with a random deviation, but most likely with a magic trick.

Theory could perhaps help to identify to what extent an observed piece of evidence is random or genuine. There is a large difference between the dice example and the research in economics, in that we actually have good reasons to believe in statistical significant relationships in economics, but we have no such reasons in case of the dice example. They have, at the start, different plausibilities. A good starting point would be to think about families of tests as sets of tests with the same plausibilities.

However, there are challenges. How exactly could we analyse cases in which many different hypotheses with different plausibilities are tested? To stick with the dice example; say, 100 magicians in training participate in a dice throwing experiment in which they may show their cheating skills. The 100 magicians all had a different number of days of training as a magicians (from 1 to 100 days). Ten of them throw double sixes (while 2.78 are expected to do so). Do we conclude that all these ten were genuine cheaters? Or, were 7 of them cheaters, and 3 of them the result of search effects? Hence, theory may help to analyse the data, but it does not solve the problem completely. Perhaps a more comprehensive solution in this area could be sought: a good point of departure for future research.

However, even if we could develop a conceptual meaningful way to define families of statistical tests, such that we can account for what is plausibly random, and what is plausibly genuine, there is still an informational problem. We still would need to know how many hypothesis tests are conducted that are relevant in the evaluation of my hypothesis. And even if we know would know what exactly this would boil down to, it would still require a lot of

information. Collecting this information may be possible for some sciences (see post-script of this chapter), but not for economics.

4.7. Conclusion: The fast and the spurious

A crucial problem for data mining is that there are many accidental correlations in the world, and searching for them makes it likely that rather than genuine correlations, accidental correlations are found. This is called the pretest bias and searching in the data is its cause. Because of this problem, the classical interpretation of statistics assumes that a model on which inferences are made is estimated only once, and if it is estimated more often, this needs to be corrected (**CCT**).

In this chapter I examined how reporting results could possibly help to identify cases that are vulnerable to pretest bias. Several authors have already questioned this possibility. I found that in observational data analysis, the amount of search that is conducted is not limited to the search conducted by a single researcher on a single research project. We do not know very well how we could possibly quantify the search behind our findings, but we do know that simply reporting and correcting for all tests conducted in the research for one paper is not enough. The publication bias is one example of a search procedure that is hard to quantify. These findings support **IT**

Consequently, the problem of pretest bias is not limited to researchers who fail to report all their search procedures, it is simply a feature of economic research. We cannot say that data mining is bad because it causes a pretest problem. If this would be so, we would have to condemn all economic research practice. As Halbert White (2000) argues: data mining, as seen as pretest bias, is "endemic".

I started the chapter by explaining the strong normative attitudes that people hold towards data mining and multiple testing. It has been called sinning to not report all the results that were conducted for the research. However, as it turns out, even if everyone would report their results, it would not be clear how this would avoid pretest bias. Reporting the results would not solve the methodological problem. The normative attitude towards it therefore seems misguided. The problem is not that researchers do not report their results. The problem is that it is unclear which results should be reported and which ones do not need to be reported, and what to do with all the reported results.

What kind of consequences does this have? **IT** may sound very pessimistic. However, remember that what is inevitable is the pretest bias, which is a problem for the inferential interpretation of test statistics. The consequence of the inevitability is simply this: an observed p-value that is very low, does not mean that the chance it was observed given the null hypothesis is low. The chance to find a low p-value may be high due to the large amount of search conducted[23]. In other words, the p-values cannot be interpreted as measure of epistemic warrant (cf. Hoover and Perez, 2000). This goes against the classical interpretation of these statistics, and it raises a new question: if this is not the way to learn from statistics,

[23] I formulate this point, again, in terms of classical statistics for rhetorical reasons. However, the same applies, to some extent to the Bayesian. If one is looking for a low likelihood, one will find low likelihoods. And if many people are gathering evidence, some misleading likelihood ratios will be expected to be found.

what is? In the following chapter I discuss to what extent data mining methods can be seen as alternative methods of inference to the classical interpretation of p-values.

On a final note, I want to point out that there is something almost paradoxical about multiple testing as seen as a methodological problem. If the problem is taken very seriously, the developments in computer power technology of the past decades has not made our results more reliable, but instead made them more likely to be spurious. The classical interpretation of statistics is a valid method of inference only if there is no search involved, which is exactly what the fast computer power allows us to do. If **CCT** is taken seriously, this may be a consequence of easy access to computer power. Should this then be taken as a plea to make computers in economic departments slower and data harder to find? That would be absurd. It is simply a plea not to interpret statistics in the classical sense in cases where one suspects much search has been conducted. However, there are alternative interpretations available, and these should be pursuit and developed further.

Interlude to chapter 4:
Lessons from the Higgs Boson Discovery

In this chapter I have tried to explain difficulties of applying the orthodox multiple testing correction framework on economic problems. An important example in which this multiple testing correction framework was applied (arguably very successfully) is the recent case of the Higgs boson discovery. Put very simply, particle physicists were working with the Standard Model which predicts a number of observations for a collision at a certain energy level. However, at some point (a so-called 'bin') in the energy level, a deviation from the model was expected in the alternative theory of Higgs. There were many candidate bins where this deviation might occur, and theory was silent on the question in which specific bin the deviation was expected. At the same time, the data were observed with random error. Consequently, deviations were observed often, and the more deviations were examined, the more significant deviations popped up. Louis Lyons, one of the statisticians at the project writes: "we have all too often seen interesting effects at the 3σ or 4σ level go away as more data are collected." (p.16).

Consequently, the particle physicists tried to correct for the multiple tests conducted. They did so quite rigorously. Because of the search conducted in the project, Lyons writes: "Thus the chance of a 5% fluctuation occurring somewhere in the data is much larger than might at first appear." (p.17). They call this problem the "Look Elsewhere Effect". In order to control for it, they tried to correct for all the searches that were conducted in the project. Trying to do this for all the researchers did turn out to be challenging, but it was indeed done.

The correction for the look elsewhere effect can be taken as a good example of how correction methods are ideally applied to statistical inference in light of multiple testing. It also illustrates some of the difficulties. In economics it would be extremely complicated to get all the researchers working on the same questions to cooperate to collect all the search procedures that are conducted. Such cooperation is simply not feasible. Moreover, in a footnote to the explanation on the look elsewhere effect, Lyons also highlights some of the

conceptual difficulties that relate to some of the problems that I have discussed in case of economics. It is insightful to quote him at length:

"The extent to which other people's searches should be included in an allowance for the 'look elsewhere' effect depends subtly on the implied question being addressed. Thus are we considering the chance of obtaining a statistical fluctuation in any of the analyses we have performed; or by anyone analysing data in our experiment; or by any Particle Physicist this year? Anyone observing a possible Higgs signal at the LHC would be very unhappy about having to reduce the significance of their result because of the statistical fluctuations that could occur in speculative searches performed elsewhere." (p.17, footnote)

If these issues are not unproblematic in the case of LHC project, they are surely quite problematic in economics, in which there is little cooperation, much more different kinds of questions being examined, and many overlapping samples being studied. The Look Elsewhere Effect may have been successfully dealt with in the example of the Higgs boson discovery, but to achieve this level of rigour if it comes multiple testing correction in economics is simply impossible. If a project like the LHC can only correct for multiple testing with much effort, economics will have no hope to do the same.

Chapter 5:
What can we learn from (mined) statistics?

" Because of the things we don't know we don't know, the future is largely unpredictable."
Singer, (1997, p. 39)

5.1. Introduction

In the former chapter I defended the *Inevitability Thesis* (**IT**). I concluded that pretest bias is inevitable, as it is conceptually unclear, and practically impossible, to quantify the amount of search relevant to the evaluation of a certain hypothesis. This thesis has some important consequences. If a low p-value should not be interpreted as a good reason to believe that a certain null hypothesis is true, what is it that we can learn from it? In this chapter I discuss in detail the way in which the pretest bias uncertainty affects statistical reasoning, and what kinds of possible ways, in light of this argument, still exist to justify beliefs on the basis of statistics. I also discuss the problems with these justifications, and the broader question how we should see data mined results.

In order to do so I first discuss three different positions one can maintain with respect to statistical reasoning in light of **IT**. Secondly, I discuss what we can learn from the automatic data mining methods discussed in chapter 2 without using a classical inference scheme. Automatic data mining methods do not use low p-values as evidence by themselves but evaluate them in the context of the procedure. These methods have been shown to be reliable in simulation experiments. However, these simulation experiments are plausibly very different from the target: economic phenomena "in the wild". I discuss to what extent we can be sure that these methods are reliable ways to learn about economic phenomena.

5.2. Three plausible positions

In order to clarify some positions in the debate, we can formulate three statements which cannot all be true at the same time. We can use them to describe a number of possible positions one can hold with respect to classical inference in light of **IT**.

(1) Valid classical statistical inference requires a sound procedure for accounting for pretest bias.

(2) Classical statistical reasoning in econometrics is valid.

(3) A sound procedure for accounting for pretest bias does not exist for econometric inference.

Pretest bias is not inevitable

Firstly, the people who do not accept my argument made in the preceding chapter may think that it is perfectly possible to quantify search and correct for it in the classical

framework. They will reject (3) but maintain (1) and (2). While I have not encountered an explicit account that defended this, Mayer (2000) does work under the assumption that correction is, at least to a large degree, possible. Moreover, most applied econometricians and textbooks work on that assumption.

Classical inference is invalid in economics

Greene (2000) argues for a thesis that is similar to (3), at least for the case of most economic practices. He accepts (1), and rejects (2), in at least many applied cases. The only way to overcome the problem according to Greene is by means of non-observational studies, such as experiments, or, if observational studies are conducted, to test hypotheses only by means of out-of-sample evidence[24]. Like Greene, Hoover and Perez (2000) discuss what I have called **CCT** and reject that correction is a way to save the epistemic interpretation of test statistics[25]. They embrace non-inferential interpretations of test statistics. Hence, Hoover and Perez (2000) would also reject (2) and accept (1) and (3).

Classical inference is still valid, because it does not require correction

Alternatively, Hurlbert and Lamberdi argue that the classical framework is perfectly useful, but should simply not be based on trying to estimate the true value of the probability evidence occurred given a hypothesis (P(E|H)). Due to similar problems I discussed, they regard knowledge of such a value out of our reach. However, they still think the classical framework for hypotheses testing is very useful. Hence, they reject (1) and accept (2) and (3).

The last two positions seem reasonable to me. If the argument from the former chapter is accepted, it is not yet clear whether (1) or (2) is the premise that should go. Those who reject premise (1), do accept that a good method to quantify search is required for the inferential interpretation of test statistics. The disagreement ends up being about whether interpreting p-values as evidence is a crucial part of classical inference. Both agree that there is something particularly wrong with interpreting statistics as evidence without considering their history.

To explain why, consider a thought experiment from Hoover (1995): a twin paradox. The first brother is a theory minded researcher who has a good theory to describe the data. He tests it on the data and he finds good evidence for his theory. The second brother is a data miner and searches the data, but ends up with the same regression. Hoover wonders if the found evidence by the second brother makes the hypothesis either more, or less, plausible? The found p-value of the second brother cannot be interpreted inferentially. But it is the same as the one that the first brother found. What it does show in both cases is that the data fits the hypothesis very well in both cases. However, in order to understand what it teaches us, we need to know more about how this data was brought about. The p-value found by the first

[24] That is, splitting up data sets in cohorts. One is used for estimation; one is used to test predictions on. The latter cohort is the out-of-sample cohort,.

[25] To be precise. They characterise the view as: The (…) attitude is the one that that we believe is the most common in the profession namely, data mining is to be avoided and, if it is engaged in, we must adjust our statistical inferences to account for it"

brother is better evidence for the hypothesis than the same value found by the second brother. Hence, p-values cannot by themselves provide epistemic warrant.

While Hurlbert and Lamberdi (2012) defend the classical framework, they share a similar view when it comes to the inferential interpretation of p-values. They argue that correction for multiple testing is not necessary. However, they argue that "SWERPs and FDRPs collectively represent historically understandable but logically unjustifiable extensions of the outdated paleo-Fisherian and Neyman–Pearsonian frameworks for significance testing"[26] (p. 36). They also propose a revision of the interpretation of classical statistics. What they refer to as the "outdated paleo-Fisherian and Neyman-Pearsonian framework for significance testing" is exactly the kind of classical reasoning that interprets all p-values as evidence. They also propose to interpret these statistics differently.

The crucial part of the problem is that the interpretation of the p-values (and other classical test statistics) and their epistemic value rely heavily on the procedure with which they were brought about. If there was indeed a selective data mining procedure which ended up with a test with a particularly low p-value, the interpretation should be very different than the one performed by someone who has run an experiment and only tested one hypothesis. Like the twin paradox exemplifies, the meaning of p-values is different. However, according to Hoover and Perez, the classical test statistics can in fact still be interpreted in a different way: simply as measures of the sampling distribution, or put slightly differently, measures of how well hypotheses correspond to the data sample.

The classical interpretation of statistics is easily disrupted. However, the question whether data mining is a methodological problem is about more than just whether it distorts classical statistical inference or not. In the earlier chapters (particularly Chapter 2) I discussed a number of different methods to mine data systematically all of which had different ways to determine whether something was evidence than simply looking at whether a p-value was lower than 5% or not.

The discussion in the following section highlights some alternative interpretations of statistics that show how one can learn from statistics without the inferential interpretation of p-values. In particular the *Gets* approach has been widely discussed by methodologists. In the following section I discuss how their approach can teach us something about how reliable mined evidence really is.

5.3. Can *Gets* get it?[27]

In order to see what we can really learn from *Gets* it is important to examine the method in detail. As discussed in Chapter 2, there exist some very impressive simulation experiments that support *Gets* as an effective data mining method. Hoover and Perez (1999; 2004) and Hendry and Krolzig (2003) provide a large variety of experiments on simulated data that show that the approach performs much better than applying standard hypothesis testing to searching the data for statistical relationships (Hoover and Perez, 1999; Hendry and Krolzig, 2003) and better than alternative data mining methods (Hoover and Perez, 2004). We

[26] SWERP's and FDRP's stand for setwise error rate probabilities and False Discovery Rate Probabilities.
[27] Credit for this title is due to Conrad Heilmann.

can take this as evidence that *Gets* is a very reliable method. The size, that is, the realized chance of making type I errors, can be estimated to be very close to the "nominal size": 5%. In fact, Hendry and Krolzig (2003) find that after some fine tuning of the algorithm, a size even smaller than 5% was achieved, without losing much power. It shows that at least under some conditions, *Gets* can get it right, that is, *Gets* can be considered a reliable method of inference. What is the reason it performs so well? And under which conditions can we expect that it performs well? Arguments supporting the reliability of the approach and counterargument of the approach will be discussed. This will teach us something about the strengths and weaknesses of the approach, and may help us explain why the method performed well in experiments and clarify whether it will be as successful if applied to studying real economic phenomena. I must note that the *Gets* method is based on the model philosophy of the LSE approach. This philosophy views econometric models as reductions from a large data generating process (DGP). In the *Gets* approach, this takes the form of a very large regression that is reduced on the basis of an automatic algorithm. I review this econometric method and look at the evidence for *Gets*.

Key in the application of the philosophy to automatic search, first developed by Hoover and Perez (1999) and further developed by Hendry and krolzig (1999; 2003) is to start from a very general model: the general unrestricted model (GUM). The GUM contains all the potential explanatory factors of the dependent variable that one wants to consider. The GUM is tested for being congruent: explaining all the crucial characteristics of the data. *Congruence* is an important concept for Hendry's methodology. It is not based on model fit only, but also on other characteristics. Another important concept that was already briefly discussed in chapter 2 is the concept of reduction: reducing the model in such a way that it still explains all the crucial characteristics of the data the GUM explained. This is done on the basis of t-tests, but at every stage of the reduction process the model is also being tested for being congruent. A final important concept that Hendry developed is the concept of *encompassing* (e.g., Hendry and Richard, 1982; Hendry, 1988). Good econometric models should be able to explain at least the same things as their rival models (and preferably more). A model that can explain the same characteristics as other models encompasses that model. Again, explaining in this sense means more than simply fitting the data better. A better fitting model may not explain all the phenomena that the other model explained. In these cases, the better fitting model does not encompass (Hendry, 1988).

What is crucial in the reduction from the general to the smaller models is that at each step a battery of diagnostics tests are run in order to check whether the model is indeed still congruent: explaining all the characteristics of the data the GUM explained (Hendry and Krolzig, 2003). For instance, a Chow structural break test is run in order to make sure that it is not the case that model only works in some subset of the data while it fails to explain others. A test to see if the normality assumption of the error terms still holds is part of the algorithm too. And there are also a number of tests run to examine if there is no residual autocorrelation: unexplained patterns in the data. Hoover (1995) stresses the importance of these tests. A spurious result is only an apparent piece of evidence because of its p-value. A good way to distinguish between genuine and spurious statistical results is to do more statistical tests. If the removal of a variable suddenly results in unexplained patterns in the data, or non-normally distributed errors, it is taken as a sign that the model does not explain the same phenomena

anymore (Hoover, 1995; Spanos, 2000). Some may worry that this makes the evidence from the algorithm less reliable as there is a higher chance that at least one test wrongly rejects, and thereby a mistake is made in the analysis. However, these diagnostic tests, or mis-specification tests, are intended to distinguish the spurious from the genuine statistical relationships and help the analysis a great deal (Hendry, 2002). These diagnostic tests are crucial for the performance of *Gets*.

P-values are used in the *Gets* method. However, they do not signify the plausibility of observing the data were the used model the true model, or anything in this regard. Instead they signify whether or not the variance in the sampling distribution is described by the model. An insignificant p-value is considered as evidence that a variable is likely not very important. However, this is taken to be the case only if the removal of the variable does not lead to a model that explains the structure of the data less well. Hence, variables that do not appear to have an effect by themselves, because they have a low p-value, may be retained in the model if they do contribute to the overall ability of the model to describe the data. False models may still have a good fit with the data, but only the most accurate models are expected to be valid reductions of the general model. Hoover and Perez (1999) refer to their search methodology as Darwinian, because it is a search mechanism designed to be a stringent test in which the most accurate model, or the fittest model, is most likely to survive. Thus, while many p-values are still used in the *Gets* method, they are not used to signify anything in themselves about the plausibility of the model. They only have meaning in the procedure by which they are produced. That is, they are evaluated with respect to the model in which they are in, and the extent to which their removal affects the congruence of the model. Hence, the inference of *Gets* does not rely on the inferential interpretation of p-values. As Hoover and Perez (1999) put it: "The evidence of strength is not found in the t-statistics[28], but in the fact of the Darwinian of the searched specifications against alternatives and in its natural relationship to the general specification" (p. 189).

Both Hoover and Perez (1999: 2000: 2004) and Hendry and Krolzig (e.g., 2003, 2004) defend the approach from different angles. An important argument (summarized in Hansen, 1996) is that the tests used for the determination of whether or not variables should be included or not (p-values of the involved coefficients and misspecification tests) are not the same as the tests used for the evaluation of the model (model fit, congruence and ecompassing). This gives a certain neutrality to the method: it is no automatic maximization of the evaluation criteria that will lead to its good statistical performance. In this sense, the risk of overfitting is reduced. A second argument is discussed in Hoover and Perez (1999; 2000) and is more abstract. They cite a mathematical result by White (1990) that shows that under the right battery of tests, the chance of finding the correct model by means of *Gets* goes to unity if the amount of available data goes to infinity. Hoover and Perez (1999) do admit that there are some serious drawbacks to this argument. While it may show that there is some hope for this approach to do well in large samples, in many cases, such as the chapter 2 case study, this does not help much. Instead, Hoover and Perez (1999) present their experiment to defend their approach.

[28] The statistics that underlie p-values

The simulation experiments remain one of the most important arguments for the approach (Hoover and Perez, 1999; 2004; Krolzig and Hendry, 2001; Hendry and Krolzig, 2003). These experiments show that *Gets* is indeed capable to draw the right inferences. However, performing well in an experimental setting does not mean that the methods will perform well outside of the experimental setting. Aris Spanos doubts that inference in simulation says too much about inference in "the wild": "The Monte Carlo simulations [from Hoover and Perez, 1999], assuming their design is not at fault, ensure the statistical adequacy of all the statistical models; a highly unlikely scenario for such a data mining activity when modelling with real data." (2000, p. 249). He argues that it is unsurprising that the *Gets* methodology works well, under the condition of the simulation environment. In that setting researchers know that there is a true model in the data set, there are no omitted variables, no errors in the data and the functional forms are known. Whether these conditions also apply to real world settings is doubtful. I fully agree with Spanos on this point, the simulation argument for the experiments deals with an external validity problem: does good performance under the experimental conditions teach us anything about the performance of the methods under different conditions?

One problem with the extrapolation of simulation results to more realistic economic settings has been made by a number of critics of the LSE approach is that if the true model is not contained in the GUM, the *Gets* algorithm will, quite obviously, not end up with the true model (discussed in Hoover and Perez, 2004). The success in the experiments is thus conditional on the GUM containing all the relevant variables. As this is difficult to evaluate, the efficacy of *Gets* is hard to assess. In the simulation environment we can simply work under the assumption that there exist a fixed process generating the data that exclusively contains a subset of variables in the data set. In reality, we never know if our list of potentially important variables indeed contains all the true determinants. Hoover and Perez (2004) do acknowledge that the observed success of the method in the simulation experiments is conditional on having all relevant variables in the GUM. However, they still take it to be evidence vis-à-vis alternative methods. Hendry and Krolzig (2003) and Hoover and Perez (1999) are much less careful in phrasing their conclusions on the basis of the simulation experiments.

This is further complicated in case of measurement errors. Consider one of Keynes critiques on Tinbergen that has so far been left undiscussed (1939). Keynes argues that Tinbergen's model could not possibly represent the world truthfully as there are many important variables that cannot be properly measured. The model thus necessarily omits important factors from the model, which affects the accuracy of the model to represent economic reality. Keynes argued that for this reason, any political, psychological and social factors cannot be considered in the analysis, as no satisfactory measures for social and political factors exist. While this may be putting it very strongly, many economically relevant variables are difficult to measure. Consider institutional variables, which are suspected to be of crucial importance for economic development (e.g. Acemoglu et al., 2001). The extent to which these issues can be measured is limited. Acemoglu et al. (2001) do this by means of doubtful proxies such as indices that represent expropriation risk. Absence of good measures of institutions in GUM will result in a false model if institutions are indeed a determining factor of growth. This may have an important impact on quality of the model. Without this

variable a very different final model will be brought about. Hoover and Perez (2004) write that if there is a true model to be discovered, then the *Gets* algorithm will select it. This is a very big "if" considering the likelihood of mismeasurement, and imperfect measures, of important factors. This problem is completely left out of the simulation experiments. In the experimental setting, there is no measurement error whatsoever; the relevant variables are exactly represented in the data as they are in the model. This is likely very different from actual statistical inference. The true data generating process may exist, but it is unlikely that we can quantify all the relevant factors. Not only is *Gets*' success conditional on the GUM containing all the relevant variables, but also on there not being any mismeasurement. Hoover and Perez (2004) write: "If anyone seriously argues that an important variable has been omitted from the specification, the appropriate response is to add that variable to the search universe and, then, to rerun the search." (p. 790). However, this may not always be an option. In case it is not, we have good reasons to doubt the final model, but we cannot fix it.

In the recent literature on *Gets,* some work has been done to incorporate the possibility that some of the variables under consideration relate to the dependent variable in a nonlinear way (Castle and Hendry, 2012). In order to do so, they add a diagnostic test that examines the presence of nonlinear behaviour of variables. Even though Castle and Hendry are hopeful about their abilities to deal with this issue in a satisfactory way, they acknowledge that automatic data mining is much more difficult in presence of nonlinearities. It is unlikely that all economic relationships which we would like to study by means of econometrics are in linear form. At least it is not something that we can know a priori. The *Gets* algorithm used for the experiments by Hoover and Perez (1999; 2004) and Hendry and Krolzig (2003) neither included nonlinearity tests, nor did it allow for nonlinear forms. These complications were left out of the equation. This means that the reliability of the extrapolation of the experimental results to real world data is conditional on there not being important nonlinearities in the model. Given the fact that we do not know how many of the relationships we study are linear, it is doubtful that the attractive results from the experiments extrapolate to research outside of the simulation environment.

It is difficult for the *Gets* approach to deal with the kind of uncertainty about missing variables, mismeasurement, and nonlinear form, because the defence of the method has mostly focused on the ideal situation: if no complications arise, the *Gets* search strategy has very nice properties. However, if not, we do not know what the properties really are. So, even if we would know that there is a good chance there is an important variable missing from the GUM, it is hard to evaluate how this should affect inference. Some of the challenges faced by *Gets* may in the future be solved in the LSE methodology spirit: test, test and test. For instance, work on nonlinearities that Castle and Hendry (2012) are conducting may result in promising ways to detect nonlinearities, that will make the method more reliable. Hendry and Mizon (2000; quoted in Hendry and Krolzig, 2003) write that they expect the *Gets* method to develop like chess computers. The first may be unsuccessful, but as they will become much better. And quite plausibly, the *Gets* method will improve to deal with some of the issues discussed here. However, some of the issues, such as the fact that there are always relevant immeasurable variables, are not easily solved.

Moreover, developing a reliable methodology is complex, and there is almost a certain circularity to it: in order to see if *Gets* is a reliable method of inference, we need to simulate

the target as closely as possible, and in order to learn what the properties of the target are, we need a method of inference. This is, perhaps, the most tricky part of methodology: we already need to know things about how the world really is, before we can say what is the best way to study it. We can often assess in very general terms the kinds of structure of reality we are dealing with. For instance, we assume that structural breaks and nonlinearities may occur in economics, which may help us to take account of them. However, it remains difficult to assess in a precise way how reliable a method of inference will be in reality, as we do not know the extent of the presence of structural breaks and nonlinearities that will affect the reliability of the method. By stepping away from the relatively simple practice of hypothesis testing, this poses a problem for evaluating the inference in *Gets*. This critique is unfair in that it is acknowledged by Hoover and Perez (2004) that the simulation methods only serve as a critical test that can reject the methods, but not show that they in fact will work. Passing the test is no proof that the method is successful, only that it is not unsuccessful. Acknowledging this (as done by Hoover and Perez, 2004; but not by Hendry and Krolzig, 2003; or Hoover and Perez; 1999) has to lead to the conclusion that the simulations really do not tell us too much about the efficacy of *Gets*, except that it is not necessarily unsuccessful.

5.4. How about the other methods?

What do the Extreme Bound Analysis (**EBA**) and Sala-i-Martin's model averaging method teach us about what we should believe? Is a robust variable one in which we should have confidence? To some extent this question is harder to answer than the same question about *Gets*. We have little evidence, save Hoover and Perez (2004), about the efficacy of these methods. While simulation methods may not show that a method is successful, it can show that it is unsuccessful, and in a way this is what Hoover and Perez (2004) do with regard to these methods. From Hoover and Perez' experiment we learn that the original **EBA** tends to have a very small size, but very little power. In other words, it does not detect the correct variables well, but the ones that are selected quite reliably are a part of the true model. The opposite applies to Sala-i-Martin's method. It has a high size, but high power too. So, it makes many mistakes in the selection of relevant variables, but the relevant variables are indeed likely to be among the selected variables. Both methods appear to be unsuccessful. However, this is partly a value judgment. There may be cases in which one wants to reduce the amount of variables, without being absolutely sure that all the selected variables are relevant. Or, one may want to find one variable about which one is absolutely sure to have a significant relationship with growth, given a large set of covariates. In these cases, Sala-i-Martin's and **EBA**, respectively, may be desirable methods of inference. The question remains: how reliable is knowledge about their size and power exactly?

The same problems that apply to *Gets* also apply to these approaches. Mismeasurement, not including the right variables in the set of potential regressors, and nonlinearities are as much a problem for these methods as they are for *Gets*. Additionally, these methods have to deal with a lack of success in simulation experiments. Moreover, in both approaches the aggregated p-values are used to determine whether a variable is robust,

which is also the main instrument for inference. Hence, the variety of tests that are used to insure the reliability of *Gets* is another strong advantage of *Gets* vis-à-vis these methods.

One particular problem related to Sala-i-Martin's model average method and **EBA** is that the inference of robustness is made conditional on a set of models that are used for the analysis. These models may contain a large variety of different kinds of variables. A variable that is correlated strongly with another variable in the model is likely not to be robust. This is due to multicollinearity: two variables that explain the same phenomena in the data crowd each other out. Hence, one model, with one variable that correlates highly to the variable of interest, may already result in the variable turning out to be insignificant even though it is in fact an important variable. Hence, our knowledge of the size of this method based on the simulation experiments is highly conditional on how closely the tested variables all relate to each other. In some applications of the method (such as Hegre and Sambanis, 2006[29]), ways have been developed to avoid highly correlated variables in the same model. This may help to reduce the dependency of the model on the problem of **MC**. But this comes at at the cost of having an incomplete robustness analysis. Moreover, in the method tested by Hoover and Perez (2004) this was not yet the case, and knowledge of how reliable these methods really are remains uncertain. Specific to BMA in contrast to **EBA** is the averaging of the model statistics (Sala-i-Martin, 1997, Fernandez et al. 2001; Sala-i-Martin et al. 2004). This may make the problem less drastic. One model in which high multicollinearity distorts the analysis can be averaged out by other models. However, the crux of the problems remains. Inference is based on the selection of models in which multicollinearity may be the most important determining factor in assigning robustness to different variables. And whether a variable is collinear with another variable has little to do with its relevance to explain a third variable, such as growth.

While this is a crucial problem to **EBA** and Sala-i-Martin's method, the same problem may also distort the analysis in the *Gets* methodology (Hollanders, 2011). A variable that is highly collinear with another variable is more likely to be insignificant and to be removed for that reason. However, because in the *Gets* algorithm the congruence of the model is checked every time a variable is removed, it is much less likely that a variable gets removed merely for its collinearity with another. After all, if the removed variable is in fact relevant to explain the variable of interest, it will be likely that the congruence of the model is worse if the variable is removed. Hence, the problem of multicollinearity is most severe for the **EBA** and Sala-i-Martin's method than it is for *Gets*.

5.5. Some concluding remarks about uncertainty

The real problem is not so much that the data mining methods are unreliable. The simulation results show, in the least, that they are reliable under certain conditions – even more reliable than classical inference if we believe Hendry and Krolzig (2003). However, the real problem is that there is no way to know if they perform this well outside of the simulation environment, when the simulations are less than perfect representations of their target systems. The arguments from the simulation results are all conditional on economic reality

[29] They use Sala-i-Martin's method, calling it sensitivity analysis, but for this specific point that is not relevant.

being like the simulations. However, a number of conditions have been assumed to be known in case of the simulations that are uncertain in realistic settings. To summarize them: 1) uncertainty of not having the true model in the list of considered variables. Mismeasurement worsens the problem, as important variables may be impossible to measure accurately. There is nothing in the *Gets* methodology, or any other methodology, that makes it likely that a wrong model will be selected in these cases. There is 2) uncertainty of functional form (the world not being linear, while the methodology models the world in a linear way). Furthermore, there is 3) uncertainty about the extent to which multicollinearity affects the reliability of the methods, which is especially a problem for **EBA** and Sala-i-Martin's method. The simulations tried to offer some insight in the methodological qualities of these methods, but in reality this does not only depend on the methods themselves, but also on how the world is. If the world is not as simple as the simulated data sets, we cannot estimate the probability of retaining a false variable in the model very well. Or, the probability of not including a relevant variable. It therefore becomes terribly difficult to critically assess the quality of the discussed data mining methods.

Hoover and Perez (2004) humbly admit this: "We can never guarantee that the specifications selected by the general-to-specific approach are true." (p. 790). We therefore have to conclude that the simulation results provide no presumption about its efficacy outside of the simulation environment. We are thus left in the dark about the reliability of the methods if we want to apply them to economic phenomena of interest. This is, I take it, the great disadvantage of the data mining methods. We do not have good arguments to say that they will not work, but we can also not say that we know that they will.

An important consequence of the uncertainty involved in the data mining methods is that inferences are not objective. Because we cannot be sure if the success in simulations can be extrapolated to other situations, we have to judge this ourselves. For instance, we may have to evaluate if it is likely in a certain context that there are important variables missing from the analysis. There are no objective methods to determine if this is the case or not. The inferential interpretation of p-values, discussed in the former chapter has appeared to have the advantage of being objective. However, the objectivity appeared to be an illusion when it is unclear to what extent a result was brought about by the search effect. As there is no objective way to account for search either, the reliability of the inferential interpretation of p-values is not objective anymore either. In light of this analysis we need to conclude that data subjectivity is a fundamental part of econometric analysis.

Chapter 6:
Summary and Conclusion

David Hendry (1980) wondered if econometrics is science or alchemy. He argued that it could be both. It is possible, he claimed, to get any result a researcher wants to have if she searches hard enough. This will not at all be reliable evidence, and the term alchemy seems well suited for this practice. Ronald Coase is credited with a famous dictum about this kind of research: "if you torture the data long enough, Nature will confess". Strong data mining is one of the most obvious ways in which the data can be tortured. At the same time, if a researcher carefully applies the three golden rules[30], he may in fact end up with very reliable results. These rules make the difference between econometrics as alchemy or econometrics as science. The general-to-specific approach is a part of Hendry's answer: econometrics can be scientific if we think hard and carefully about methodology. His approach has been widely discussed in this thesis and in summarizing the conclusions from my thesis I discuss whether Hendry's method is indeed to be called science rather than alchemy.

The most general view about data mining is that it is bad practice. I have called this view the *Condemnation Thesis*. At the same time it has been defended as a desirable practice, a widespread practice, an inevitable practice and a warranted practice. These claims have been critically discussed. Firstly, in Chapter 2 I described the claim that data mining is in fact something that is very desirable in the field of growth economics. There are certain economic problems that simply go unsolved without data mining, in particular when theory cannot solve the problem for us. However, a desire for gold does not make alchemy a proper scientific practice. Therefore, I dealt with the main reasons why data mining is considered problematic in Chapter 3 and 4 and showed that they are not very good reasons to condemn data mining altogether.

The problem discussed in Chapter 3 was double counting. A researcher who uses the data both in the construction of hypotheses and the testing thereof seems to do something very unscientific. This is both an intuitive and a widely accepted view of science. However, I defended that this is not a problem in itself. The true problem is the arbitrariness that people associate with post hoc inference. Hendry's methodology is specifically designed to get the most reliable evidence as possible, and every test conducted in his methodology is intended to avoid arbitrariness. Its application in an automatic computer algorithm cannot be used to arrive at any desired result. The method is neutral to the desired outcome. Moreover, avoiding post hoc inference in search algorithms may come at a high cost for other virtues of evidence such as precision. The accusation of double counting is therefore not one that should keep us from seeing Hendry's method as scientific. So far, it passes the science test.

The fourth and fifth chapter dealt with the statistical inference. Perhaps the most problematic consequence of data mining is that it distorts the inferential interpretation of statistical evidence. Looking for statistically significant evidence makes it likely evidence is found even when there is in fact no real deviation from the hypotheses due the search effect.

[30] That is: test, test and test.

In chapter 4 I argued that the assumption that a statistical model is only estimated once is quite strange to begin with. Especially now that regressions only take milliseconds to compute, this assumption seems unrealistic and undesirable. A common view, as is expressed by Hollanders (2011) or Mayer (1980; 1992; 2000) is that we can either work under this assumption, or we need to keep track of the search that was conducted in order to correct for it. I called this view the *Corrective Condemnation Thesis*. However, keeping track of search conducted to arrive at a hypothesis is by no means a realistic procedure. Keeping track of the statistical inferences in a way that is required for multiple testing correction would not only be practically infeasible, but conceptually unclear too. What exactly are all the relevant statistical inferences? I could not find a good answer to this question. I discussed the view that the reason this assumption is violated is due to moral hazard of econometricians, and argued that violating the assumption is inevitable. Thereby, the moral hazard is unavoidable, and therefore it is not really moral hazard. So, while Hendry's methodology in fact breaches an assumption needed for the interpretation of statistics in the classical manner, so does every other method that is used to arrive at good econometric models.

This may sound very pessimistic. However, different from text book economics, Hendry's methodology actually does not rely on the classical interpretation of these statistical tests, but develops and alternative account of learning from statistics. Hoover and Perez (2000), for instance, make this explicit. This poses a new problem. While the classical interpretation of statistics may be flawed in econometric practice, it does offer a conceptually simple method to derive evidence from statistics. If this method is dropped, we need an alternative. Attempts at such an alternative are developed by Hoover and Perez (1999; 2000; 2004) and Hendry and Krolzig (1999; 2003). The general intuition behind the general-to-specific approach is that reduction is much more likely to end up with the best model, as only the variables that do not seem to contribute are removed, and the rest of the model is retained. The most important argument in favour of the *Gets* approach is the success that they achieve in simulation experiments.

In Chapter 5 I raised many doubts about the extrapolation of these results to the real world. Hoover and Perez (2004) are modest in their claims and argue that the results of the simulations merely show that the method can perform well, if the general model indeed contains the true model. In actual econometric practice, we never know whether this is true. So, we can be sure that *Gets* can perform well, but the real problem seems to be that we do not know if it will perform well outside of the simulation setting.

The discussion in the last chapter ended on this note. And we do not seem to have a clear answer yet to the question we set out to answer: *Gets*: alchemy or science? The two main objections to data mining do not seem to show convincingly that data mining is unscientific, but the arguments in favour of the inference in *Gets* do not show that it is. The unfortunate consequence of searching for an answer to a complicated question is that the answer might be complicated too.

To make the matter more concrete recall the results that the *Gets* approach found in the growth economics debate. The model describing growth in cross-country growth regressions with which both Hoover and Perez (2004) and Hendry and Krolzig (2004) ended up contains: 1) fraction of the population Confucian, 2) number of years open economy, 3) equipment investment, 4) revolutions and coups (negative), and 5) fraction of the population

Protestant (negative). The question we need to answer is whether this is the true model describing the data generating process of growth data in the world, or whether it is an alchemistical brew that may provide statistical satisfaction but is useless in its description of the world?

I believe that the final answer should remain the reader's judgment. Arguments discussed in this thesis, though, point out *Gets* is plausibly the best available method in case of the problems present in growth economics. Consequently, the model in Hendry and Krolzig (2004) and in Hoover and Perez (2004) is the best available solution to the problem of model selection in this subfield of economics. However, I also believe that there are many reasons to be sceptical about this result. Angrist and Pischke (2010), for instance, argue that the importance of the Fraction of the Population that is Confucion in cross-country growth regressions is an argument against taking them all too seriously. I find it unlikely that it is really the religion that matters, and plausibly, it is a by-product of the many important economic developments between 1960 and 2000 that brought about economic development in Confucian countries. Probably not all of these developments were measured, or were measurable. This made it the case that the Confucian dummy is the closest measurable proxy of these developments and therefore ended up in the model. Most likely something similar occurred with the Protestant variable. The absence of good measurements of the true underlying causes of the development of Asian Tigers may also be a cause of concern for the model at large, whose development is conditional on the true model being contained in the general one. It is important to keep these issues in mind. The *Gets* approach may be a promising one, but given how difficult it is to make inferences about the world, we should be careful in interpreting its results.

The kind of judgment that needs to be made in interpreting the results from a *Gets* procedure is whether the context in which the testing occurs is similar to the simulation setting in which *Gets* has been tested. And if not, to what extent is it harmful for the result? This kind of judgment is very subjective. In terms of Hendry's terminology we might say that this makes econometrics a little alchemistic. Two different researchers may make different inferences, or draw different conclusions, because they have different views on the matter. However, we need to keep in mind that subjective judgments do not necessarily mean that the judgment is unreliable. Judgments may be guided by good reasons. The experienced econometricians would likely make these judgments much better than a college freshman in econometrics 101. This is not necessarily a problem. Reiss (2013), for instance, argues that mechanical objectivity is by no means something necessarily preferable to proper expert judgment. Many would share this view.

Consider the scope of the argument. In Chapter 1 we discussed that the definition of data mining I provided does not only include large searches, but included smaller degree searches. Given that there is multiple testing in almost every research conducted in econometrics, this point has relevance for a very large share of cases. Moreover, we also discussed unintentional data mining as a possibility and in particular in Chapter 4 we concluded that pretest bias due to search is something that is extremely difficult to avoid. The widespread nature of this problem allows us to draw some conclusions on the basis of our discussion of data mining about economic methodology at large.

The examples discussed in this thesis were almost always the obvious cases: one (group of) researcher(s) conducting a large search over a set of variables that are suspected to explain the variable of interest. In these cases, treating the results as if they do not come from a search is very obviously misleading. However, in smaller instances, or unintentional instances, this issue is somewhat less obvious even though it is equivalent in nature. The analysis in Chapter 4 and 5 suggests that the statistics from the smaller degree cases of data mining are affected by the problem of pretest bias just as well as the larger degree cases. If we consistent, we have to accept that the inferential interpretation of p-values is not only problematic in the large data mine case, but also in the smaller degree ones; that is, almost all statistical analysis in economics. We have to conclude: p-values do say much.

I am by far not the first to discuss this issue. De Long and Lang (1992) wonder "Are all economic hypotheses false?", and McCloskey and Ziliak (2004; 2008) speak of the cult of statistical significance[31]. Quite interestingly in this respect was the discussion by Keuzenkamp and Magnus (1995) who awarded a prize to someone who could provide an example of a significance test that actually changed the received view in a subfield of economics. No one won[32]. Apparently, economists themselves do not consider the hypothesis testing as a crucial argument for truth. My discussion in Chapter 4 and 5 supports p-value scepticism. The assumption that only one model is estimated and considered in order to interpret the statistical tests inferentially (or otherwise all the other results are corrected for multiple testing) makes the procedure inapplicable. Even though p-value scepticism is not original, it is still relevant as economics journals still focus much on this statistic (McCloskey and Ziliak, 2008).

What to do instead? A possible suggestion is to always use *Gets* for all empirical research in economics. However, we need to learn more about its inferential properties in imperfect situations before we can put our trust in the method. A less rigorous consequence would be to keep using the classical method of inference in economics, but to accept a truth that is likely already accepted by many applied statisticians: a p-value does not tell us anything about the strength of the evidence against the null hypothesis by itself. We simply need to understand p-values as properties of the data with respect to the hypothesis, and nothing more than that. In order for us to understand how strongly this can be taken as evidence against the hypothesis under test we need to be able to understand how search effects have influenced the discovery of the result. In many subfields of economics we may have useful intuition about this. For instance, in case of the cross-country growth regressions, we know how simple it is to estimate many models and see which ones do well. Thus, scepticism about empirical results is in place. In experimental economics, search is much less likely to have brought about results, which should provide some confidence in the reliability of the empirical results. And some instances of empirical research in economic lie in between these extremes. Consider Acemoglu et al. (2001). They conducted an instrumental variable approach to estimate the effect of institutions on growth by using settler mortality rates as an instrument for early institutions, which they believe are related to today's institutions, which they estimate by means of an index of expropriation risk. In order to arrive at this result they

[31] Admittedly, for different reasons.

[32] Personally I think that if the competition would be held today, many might suggest Acemoglu et al. (2001) as an example of a powerful significant statistical test that resulted in a strong boost for new institutional economic thinking.

had to do a lot of historical research, and go to libraries around the world to gather the data. Moreover, they found numerous pieces of alternative sources of evidence to support their significant result. In this case, one can be sure that the search selection effect is relatively small, and that the arbitrariness of this statistical result was not considerably high. Whether one can say that the probability that the result came about given the evidence truly corresponds to the reported p-values is a matter of degree.

To conclude, econometrics is not alchemy, but one has to be particularly careful with the interpretation of test statistics. From the perspective of data mining, the standard interpretation of statistics that is taught in econometrics 101 seems to be misleading in almost all cases in which econometrics can be applied. But while much work is still to be done, promising alternatives exist and the problems related to data mining come in different degrees. We do not need to worry that all econometrics is a waste of time, or completely unscientific. We merely need to be very careful in interpreting the results.

References

Austin, P.C., Mumdani, M.M., Juurlink, D.N. and Hux, J.E. (2006). Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. *Journal of Clinical Epidemiology*, Vol. 59, pp.964-969

Abdi, H. (2007). Bonferroni and Sidak corrections for multiple comparisons. In N.J. Salkind (Eds.): *Encyclopedia of measurement and statistics*. Thousand Oaks (CA): Sage. pp. 103–107.

Acemoglu, D. Johnson, S. and Robinson, J.A. (2001). The Colonial Origins of Comparative Development: An Empirical Investigation. *American Economic Review, 91(5): pp.1369-1401*

Angrist, J, and Pischke, J-S. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. *Journal of Economic Perspectives,* 24(2), pp. 3-30

Backhouse, R. E., & Morgan, M. S. (2000). Introduction: is data mining a methodological problem?. *Journal of Economic Methodology*, 7(2), 171-181.

Banerjee, A. V., & Duflo, E. (2011). *Poor economics: A radical rethinking of the way to fight global poverty*. PublicAffairs.

Barro, R. J. (1991). Economic growth in a cross section of countries. *The quarterly journal of economics*, 106(2), 407-443.

Benjamini & Hochberg, (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57, 289–300*

Brock, W. A., & N Durlauf, S. (2000). *Growth economics and reality*. National Bureau of Economic Research (No. w8041).

Burns, A. F., & Mitchell, W. C. (1946). *Measuring business cycles*. NBER Books.

Burger, R.P. & Du Plessis, S.A. (2006). *Quantifying the extent of data mining in applied econometrics*. Stellenbosch, mimeograph.

Campos Fernández, J., Ericsson, N., & Hendry, D. (2005). General-to-specific modeling: An overview and selected bibliography. *FRB International Finance Discussion Paper, (838).*

Cartwright, N. (2007). Are RCTs the Gold Standard?. *BioSocieties 2:11-20.*

Castle, J. L. (2005). Evaluating PcGets and RETINA as Automatic Model Selection Algorithms*. *Oxford Bulletin of Economics and Statistics, 67(s1), 837-880.*

Castle, J. L., & Hendry, D. F. (2012). Automatic selection for non-linear models. In Wang, L., Garnier, H. and Jackman, T. (eds.) *System Identification, Environmental Modelling and Control*, Springer, pp. 229-250

Caudill, S. (1990). The Necessity Of Mining Data. *Atlantic Economic Journal*, 16(3), pp. 11-18.

Caudill, S. (1990). Econometrics in theory and practice. *Eastern economic journal*, XVI(3). pp.249-256.

Chatfield, C. (1995). Model uncertainty, Data Mining and Statistical Inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society).* 158(3), pp. 419-466

Cuaresma, C. J., & Doppelhofer, G. (2007). Nonlinearities in cross-country growth regressions: A Bayesian averaging of thresholds (BAT) approach. *Journal of Macroeconomics*, 29(3), pp. 541-554.

De Long, J. B., & Lang, K. (1992). Are all economic hypotheses false?. *Journal of Political Economy*, pp. 1257-1272.

Deaton, A. (2010). Instruments, development, and learning about randomization. *Journal of Economic Literature*, *48*, pp. 424-455.

Denton, F.T. (1985). Data mining as an industry. *The review of economics and statistics,* 67 (1), pp. 124-127

Du Plessis, S.A. (2009). The miracle of the Septuagint and the promise of data mining in economics. In *Philosophy of* Economics (*424*-454), Oxford University Press: New York, USA.

Durlauf, S. N., Johnson, P. A., & Temple, J. R. (2005). Growth econometrics. *Handbook of economic growth*, 1, 555-677.

Durlauf, S. N., & Quah, D. T. (1999). The new empirics of economic growth. *Handbook of macroeconomics*, 1, 235-308.

Fernandez, C., Ley, E., & Steel, M. F. (2001). Model uncertainty in cross-country growth regressions. *Journal of applied Econometrics*, 16(5), 563-576.

Feyrer, J. (2009). Distance, Trade, and Income – The 1967 to 1975 Closing of the Suez Canal as a Natural Experiment. *NBER Working Papers 15557*. National Bureau of Economic Research, Inc.

Fisher, R. A. (1935). The design of experiments.London: Oliver & Boyd.

Fisher, R. A. (1956). Statistical methods and scientific inference. New York: Hafner.

Frankel, J. A. and Romer, D. (1999). Does trade cause growth? *American Economic Review*,

89(3): pp. 379-399.

Friedman, M. (1940). Review of Tinbergen (1939). *American Economic Review*, 30(3), 657-60.

Friedman, M. (1953). The methodology of positive economics. In Hausman, D. (eds). *The Philosophy of economics: an anthology*, *2*, pp. 180-213.

Gilbert, C. L. (1986). Practitioner's Corner: Professor Hendry's Econometric Methodology. *Oxford Bulletin of Economics and Statistics*, 48(3), 283-307.

Glymour, C., Madigan, D., Pregibon, D., & Smyth, P. (1997). Statistical themes and lessons for data mining. *Data mining and knowledge discovery*, 1(1), 11-28.

Greene, C. A. (2000). I am not, nor have I ever been a member of a data-mining discipline. *Journal of Economic Methodology*, 7(2), 217-230.

Gujarati, D. N.,(2003), *Basic econometrics*. New York: MeGraw-Hill.

Haavelmo, T. (1944). The probability approach in econometrics. Econometrica: *Journal of the Econometric Society*, iii-115.

Hegre, H., & Sambanis, N. (2006). Sensitivity analysis of empirical results on civil war onset. *Journal of Conflict Resolution*, 50(4), 508-535.

Hendry, D. F. (1980). Econometrics-alchemy or science?. *Economica*, pp. 387-406.

Hendry, D.F. (1988) "Encompassing," *National Institute Economic Review*, pp. 88-92.

Hendry, D. F. (2000). *Econometrics: alchemy or science?: essays in econometric methodology*. OUP Catalogue.

Hendry, D.F. (2002). Applied econometrics without sinning. *Journal of Economic Surveys.* 16, pp. 591-604.

Hendry, D. F., & Krolzig, H. M. (1999). Improving on 'Data mining reconsidered'by KD Hoover and SJ Perez. *The econometrics journal*, 2(2), 202-219.

Hendry, D.F. and Krolzig, H.-M. (2003), New Developments in Automatic General-to-specific Modelling. 379-419 in Stigum, B.P. (eds). *Econometrics and the Philosophy of Economics*, Princeton University Press.

Hendry, D. F., & Krolzig, H. M. (2004). We Ran One Regression*. *Oxford bulletin of Economics and Statistics*, 66(5), pp. 799-810.

Hendry, D., & Mizon, G. (2000). Reformulation empirical macroeconomic modelling. *Oxford Review of Economic Policy*, 16(4), pp. 138-159.

Hendry, D. F., & Morgan, M. S. (1995) eds. *The foundations of econometric analysis.* Cambridge University Press.

Hendry, D. F., & Richard, J. F. (1982). On the formulation of empirical models in dynamic econometrics. *Journal of Econometrics*, 20(1), pp. 3-33.

Hitchcock, C., & Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *The British journal for the philosophy of science*, 55(1), pp. 1-34.

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. John Wiley & Sons, Inc..

Hollanders, D.A. (2011). Five methodological fallacies in applied econometrics. *Real-world economics review*, 57, pp. 115-126

Hoover, K. D. (1995). In defense of data mining: some preliminary thoughts. *Monetarism and the Methodology of Econonomics: Essays in Honor of Thomas Mayer*, pp. 242-57.

Hoover, K. D., & Perez, S. J. (1999). Data mining reconsidered: encompassing and the general-to-specific approach to specification search. *The Econometrics Journal*, 2(2), pp. 167-191.

Hoover, K. D., & Perez, S. J. (2000). Three attitudes towards data mining. *Journal of Economic Methodology*, 7(2), pp. 195-210.

Hoover, K. D., & Perez, S. J. (2004). Truth and Robustness in Cross-country Growth Regressions*. *Oxford bulletin of Economics and Statistics*, 66(5), pp. 765-798.

Howson, C. (1988). Accommodation, prediction and bayesian confirmation theory. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* (pp. 381-392). Philosophy of Science Association.

Howson, C. (1990). Fitting Theory to the Facts: Probably Not Such a Bad Idea After All. *Scientific Theories*, Univ. Minnesota Press, Minneapolis.

Howson, C. (1991). The 'old evidence' problem. *The British Journal for the Philosophy of Science*, 42(4), pp. 547-555.

Howson, C., & Urbach, P. (2005). *Scientific reasoning: the Bayesian approach*

Hurlbert, S. H., & Lombardi, C. M. (2012). Lopsided reasoning on lopsided tests and multiple comparisons. *Australian & New Zealand Journal of Statistics*, 54(1),pp. 23-42.

Islam, N. (1995). Growth empirics: a panel data approach. *The Quarterly Journal of Economics*, 110(4), pp. 1127-1170.

Jegadeesh, N. and Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance*, 48, pp 65–91

Kaldor, N. (1957). A model of economic growth. The economic journal, 67(268), pp. 591-624.

Kennedy, P.E. (2002). Sinning in the Basesment: What are the Rules? The ten Commandments of Applied Econometrics. *Journal of Economic Surveys*, 16(4), pp. 569-589.

Keynes, J.M. (1939). Professor Tinbergen's Method. *The Economic Journal*, 49(195), pp. 558-577.

Keynes, J. M. (1940). On a method of statistical business-cycle research. A comment. *The Economic Journal*, 49(197), pp. 154-156.

Keuzenkamp, H. A. (1995). The econometrics of the Holy Grail—a review of econometrics: alchemy or science? Essays in econometric methodology. *Journal of Economic Surveys*, 9(2), pp. 233-248.

Keuzenkamp, H. A., & Magnus, J. R. (1995). On tests and significance in econometrics. *Journal of Econometrics*, 67(1), pp. 5-24.

Koopmans, T. C. (1947). Measurement without theory. *The Review of Economics and Statistics*, 29(3), pp. 161-172.

Krolzig, H. M., & Hendry, D. F. (2001). Computer automation of general-to-specific model selection procedures. *Journal of Economic Dynamics and Control*, 25(6), pp. 831-866.

Leamer, E. E. (1978). *Specification searches: ad hoc inference with nonexperimental data*. New York: Wiley.

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, 73(1), pp. 31-43.

Leamer, E. E. (1985). Sensitivity analyses would help. *The American Economic Review*, 75(3), pp. 308-313.

Leamer, E. E. (2010). Tantalus on the Road to Asymptopia. *The Journal of Economic Perspectives*, 24(2), pp. 31-46.

Levine, R., & Renelt, D. (1992). A sensitivity analysis of cross-country growth regressions. *The American economic review*, 84(4), pp. 942-963.

Ley, E., & Steel, M. F. (1999). *We Just Averaged over Two Trillion Cross-Country Growth Regressions*. International Monetary Fund.

Liew, V. K. S. (2004). Which lag length selection criteria should we employ?. *Economics Bulletin*, *3*(33), pp. 1-9.

Lo, A. W., & MacKinlay, A. C. (1990). Data-snooping biases in tests of financial asset pricing models. *Review of Financial studies*, 3(3), pp. 431-467.

Lo, A.W., & MacKinlay, A. C. (1999). *A Non-Random Walk Down Wall Street*. Princeton, Princeton, New Jersey.

Lucas, R.E., Jr. (1988). On the Mechanics of Economic Development. *Journal of Monetary Economics*, 22(1), pp. 3-42.

Lyons, L. (2008). Open statistical issues in particle physics. *The Annals of Applied Statistics*, pp. 887-915.

Magnus, J. R. (1979). Substitution between energy and non-energy inputs in the Netherlands 1950-1976. *International Economic Review*, 20(2), pp. 465-484.

Magnus, J.R. (1999). The success of econometrics. *De Economist*, 147(1), pp. 55-71.

Magnus, J.R. (2002). The missing tablet: comment on Kennedy's ten commandments. *Journal of Economic Surveys*. 16, pp. 605–609

Magnus, J. R., Powell, O., & Prüfer, P. (2010). A comparison of two model averaging techniques with an application to growth empirics. Journal of Econometrics, 154(2), pp. 139-153.

Mäki, U. (2009). Unrealistic assumptions and unnecessary confusions: Rereading and rewriting F53 as a realist statement. *The methodology of positive economics: Reflections on Milton Friedman's legacy*, pp. 90-116.

Mayer, T. (1980). Economics as a hard science: Realistic goal or wishful thinking?. *Economic Inquiry*, 18(2), pp. 165-178.

Mayer, T. (1992). *What Do Significance Tests Signify?* Department of Economics, University of California, Davis, Working Paper No. 397.

Mayer, T. (2000). Data mining: a reconsideration. *Journal of Economic Methodology*, 7(2), pp. 183-194.

Mayo, D.G. (1996), *Error and the Growth of Experimental Knowledge* (Chapters 8, 9, 10). University of Chicago Press, Chicago.

Mayo, D. G. (2008). How to discount double-counting when it counts: Some clarifications. *The British Journal for the Philosophy of Science*, 59(4), pp. 857-879.

Mayo, D. G. (2010). An Ad Hoc Save of a Theory of Adhocness? *Exchanges with John Worrall (pp. 155-169*). Cambridge: Cambridge University Press.

Mayo, D. G., & Kruse, M. (2001). Principles of inference and their consequences. In *Foundations of Bayesianism* (pp. 381-403). Springer: Netherlands.

McCloskey, D. and Ziliak, S. (2004). *Size Matters: The Standard Error of Regression in the American Economic Review*. Journal of Socio-Economics, 33, pp.527-546

McCloskey, D. and Ziliak, S. (2008). *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. University of Michigan Press

Miller, R.G., Jr. (1981). *Simultaneous Statistical Inference, 2nd*. New York: Springer.

Neuhierl, A., & Schlusche, B. (2011). Data snooping and market-timing rule performance. *Journal of Financial Econometrics*, 9(3), pp. 550-587.

Pagan, A. R., & Veall, M. R. (2000). Data mining and the econometrics industry: comments on the papers of Mayer and of Hoover and Perez. *Journal of Economic Methodology*, 7(2), pp. 211-216.

Perez-Amaral, T., Gallo, G. M., & White, H. (2003). A Flexible Tool for Model Building: the Relevant Transformation of the Inputs Network Approach (RETINA)*. *Oxford Bulletin of Economics and Statistics*, 65(1), pp. 821-838.

Perez-Amaral, T., Gallo, G. M., & White, H. (2004). A comparison of complementary automatic modeling methods: RETINA and PcGets. *Econometric Theory*, 21(1), pp. 262-277.

Reiss, J. (2008). *Error in economics: Towards a more Evidence-bas Methodology*. Abingdon: Routledge.

Reiss, J. (2013). Struggling over the Soul of Economics: Objectivity vs. Expertise. In Boumans, M. and Martini, C. (eds). *Experts and Consensus in Social Science*, New York (NY): Springer

Rodrik, D. (2009). The new development economics: we shall experiment, but how shall we learn?. In: Cohen, J. And Easterly, W. (eds.) *What Works in Development*?. Brooking Institution Press.

Romer, P.M. (1986). Increasing Returns and Long-Run Growth. *Journal of Political Economy*, 94(5), pp. 1002-37.

Romer P.M. (1994). The Origins of Endogenous Growth. *The Journal of Economic Perspectives*, 8(1), pp. 3-22

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin*, 86(3), pp. 638-641.

Saatsi, J., & Vickers, P. (2011). Miraculous success? Inconsistency and untruth in Kirchhoff's diffraction theory. *The British Journal for the Philosophy of Science*, 62(1), pp. 29-46.

Sala-i-Martin, X. X. (1997). I just ran two million regressions. *The American Economic Review*, 87(2), pp. 178-183.

Sala-i-Martin, X.X., Doppelhofer, G., & Miller, R. I. (2004). Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach. *The American Economic Review*, 94(4), pp. 813-835.

Salimans, T. (2012). Variable Selection and Functional Form Uncertainty in Cross-Country Growth Regressions. *Journal of Econometrics*, 171, pp. 267–280

Saville, D.J. (1990). Multiple comparison procedures: the practical solution. *American Statistic journal*, 44, pp. 174–180.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica: Journal of the Econometric Society*, pp. 1-48.

Sims, C. A. (2010). But economics is not an experimental science. *The Journal of Economic Perspectives*, 24(2), pp. 59-68.

Scargle, (2000). Publication Bias: The "File-Drawer" Problem in Scientific Inference. *Journal of Scientific Exploration*, 14(1), pp. 91-106

Singer, M. (1997). Thoughts of a nonmillenarian. *Bulletin of the American Academy of Arts and Sciences*, 51(2), pp. 36-51.

Solow, R. M. (1956). A contribution to the theory of economic growth. *The quarterly journal of economics*, 70(1), pp. 65-94.

Spanos, A. (2000). Revisiting Data Mining: 'hunting' with or without a license. *Journal of Economic Methodology*, 7(2), pp. 231-264.

Stanley, T. D., & Doucouliagos, H. (2010). Picture this: a simple graph that reveals much ado about research. *Journal of Economic Surveys*, 24(1), pp. 170-191.

Steele, K. (2012). Persistent Experimenters, Stopping Rules, and Statistical Inference. *Erkenntnis*, pp. 1-25.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American statistical association*, 54(285), pp. 30-34.

Stegenga, J. (2011). Is meta-analysis the platinum standard? *Studies in History and Philosophy of Biological and Biomedical Sciences*, 42(4). pp. 497-507

Sullivan, R., Timmermann, A., & White, H. (2001). Dangers of data mining: The case of calendar effects in stock returns. *Journal of Econometrics*, 105(1), pp. 249-286.

Tinbergen, J. (1939). *Statistical testing of business-cycle theories: Business cycles in the United States of America, 1919-1932 (Vol. 2)*. League of nations, Economic intelligence service.

Tullock, G. (1959). Publication decisions and tests of significance—a comment. *Journal of the American Statistical Association*, 54(287), pp. 593-593.

Ulasan, S. (2011). *Cross-Country Growth Empirics and Model Uncertainty: An Overview*. Central Bank of the Republic of Turkey Working Paper, No 11/02

Verbeek, M. (2008). *A Guide to Modern Econometrics, 3rd ed*. Chichester: John Wiley and Sons.

Wacziarg, R. (2002). Review of easterly's the elusive quest for growth. *Journal of Economic Literature*, 40(3), pp. 907-918.

Werndl, C., & Steele, K. S. (2013). Climate models, confirmation and calibration. *The British journal for the philosophy of science.* (in Press).

White, H. (1990). A consistent modelling procedure based on m-testing. In Granger, C.W.J. (eds). *Modeling Economic series: Readings in econometric methodology*. pp. 369-83. Oxford: Clarendon press.

White, H. (1998). *Artificial neural network and alternative methods for assessing naval readiness.* Technical Report, NRDA, San Diego

White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5), pp. 1097-1126.

Worrall, J. (2006). *Theory-confirmation and history* (pp. 31-61). Springer Netherlands.

Worrall, J. (2010). Theory Confirmation and Novel Evidence. In: Mayo, D.G. and Spanos, A. (eds). *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science* (pp.125-154). Cambridge: Cambridge University Press.

Wooldridge, (2009). *Introductory Econometrics: a modern approach* (4th ed.). Canada: South-Western.