SCHOOL OF ECONOMICS AND MANAGEMENT
FACULTY OF SOCIAL SCIENCES
AARHUS UNIVERSITY

CREATES
Center for Research in Econometric Analysis of Time Series

CREATES Research Paper 2009-17

Statistical vs. Economic Significance in Economics and Econometrics: Further comments on McCloskey & Ziliak

Tom Engsted

School of Economics and Management
Aarhus University
Bartholins Allé 10, Building 1322, DK-8000 Aarhus C
Denmark

Statistical vs. Economic Significance in Economics and Econometrics: Further comments on McCloskey & Ziliak*

Tom Engsted†

Abstract

I comment on the controversy between McCloskey & Ziliak and Hoover & Siegler on statistical versus economic significance, in the March 2008 issue of the *Journal of Economic Methodology*. I argue that while McCloskey & Ziliak are right in emphasizing 'real error', i.e. non-sampling error that cannot be eliminated through specification testing, they fail to acknowledge those areas in economics, e.g. rational expectations macroeconomics and asset pricing, where researchers clearly distinguish between statistical and economic significance and where statistical testing plays a relatively minor role in model evaluation. In these areas models are treated as inherently misspecified and, consequently, are evaluated empirically by other methods than statistical tests. I also criticise McCloskey & Ziliak for their strong focus on the size of parameter estimates while neglecting the important question of how to obtain reliable estimates, and I argue that significance tests are useful tools in those cases where a statistical model serves as input in the quantification of an economic model. Finally, I provide a specific example from economics - asset return predictability - where the distinction between statistical and

*Accepted for publication in *Journal of Economic Methodology*. Final version: March 4, 2009 (first version: June 30, 2008). I acknowledge support from *CREATES* (Center for Research in Econometric Analysis of Time Series), funded by the Danish National Research Foundation. I thank Steve Ziliak, Katarina Juselius, and Søren Johansen for their active participation in the *CREATES* symposium "Statistical vs. Economic Significance in Economics and Econometrics" at the University of Aarhus, June 2008, and Steve Ziliak, Deirdre McCloskey, and Kevin Hoover for subsequent e-mail conversations. I also gratefully acknowledge the comments from two anonymous referees.

†*CREATES*, School of Economics and Management, University of Aarhus, Building 1322, DK-8000 Aarhuc C., Denmark. E-mail: tengsted@creates.au.dk.

economic significance is well appreciated, but which also shows how statistical tests have contributed to our substantive economic understanding.

Keywords: Statistical and economic significance, statistical hypothesis testing, model evaluation, misspecified models.

JEL codes: B41, C10, C12

1 Introduction

The March 2008 issue of *Journal of Economic Methodology* contains an interesting discussion between Kevin Hoover and Mark Siegler on the one hand, and Deirdre McCloskey and Stephen Ziliak on the other hand, about statistical vs. economic significance in economic research (Hoover and Siegler, 2008a,b, McCloskey and Ziliak, 2008). The pivot of the discussion is McCloskey & Ziliak's long-lasting critisism of the standard practice of statistical significance testing in applied economic research and in particular their claim that the majority of applied economists and econometricians have failed - and continue to fail - in distinguishing between statistical and economic significance. Their recent book, "The Cult of Statistical Significance: How the standard error costs us jobs, justice, and lives" (Ziliak and McCloskey, 2008) summarizes their argument and illustrates the (bad, in their view) practices in various fields such as economics, psychology, and medicine. Hoover & Siegler argue, on the contrary, that there is no convincing evidence that the economics profession systematically mistake statistical significance for economic significance. Underlying the discussion is a fundamental difference in opinion on the usefulness of classical statistical hypothesis testing procedures in scientific research. Hoover & Siegler (and most other economists) find such procedures a valuable tool for scientific discovery, while McCloskey & Ziliak find them more or less useless.

In the present paper I provide further discussion of the matter. The main point I want to make is that McCloskey & Ziliak overlook important areas in economics where researchers distinguish between statistical and economic significance and where the limitations of statistical hypothesis testing are clearly acknowledged, namely those disciplines where economic models are considered to be *inherently misspecified* in the sense that there are non-negligible systematic deviations between the data and the model. Statistical hypothesis testing procedures obviously face special limitations when it comes to examination and evaluation of models with inherent misspecification. This is related to the discussion between McCloskey & Ziliak and Hoover & Siegler on sampling error (or statistical error) versus 'real error'. Hoover & Siegler

seem to confine attention to econometrically well-specified models only, i.e. models with 'well-behaved' error terms. They advance a modelling strategy (in the spirit of Haavelmo, 1944, and Hendry, 1995) in which specification testing is used to obtain errors (i.e. deviations between data and the economic model structure) that conform to a tractable probability model, c.f. Hoover and Siegler (2008a, pp.22-23). Usually within this modelling framework an important goal is to obtain errors that are completely unsystematic (identically and independently distributed, *iid*). Classical hypothesis testing naturally plays an important role in this strategy. However, an important point in McCloskey & Ziliak's discussion is that the presence of what they - with reference to W.S. Gossett, a.k.a. "Student", the inventor of the '*t*-test' - call 'real error', i.e. non-sampling error that cannot be eliminated through specification testing, is more important than sampling error, c.f. McCloskey and Ziliak (2008a, pp.41-42), and Ziliak and McCloskey (2008, pp.6-7, 16, 24, 245-246). There is general agreement in the economics profession that no economic model should be considered literally true. Models are build on simplifying assumptions, thus they are by construction only *approximations* to reality. Economists often say "models are neither true nor false", e.g. Leamer (2004). The dividing line in the profession is whether we should search for models with unsystematic *iid* errors (which is what Hoover & Siegler aim at), or instead acknowledge the inherent falseness of any model. Economists to an increasing degree hold the view that in order to be interpretable and consistent with basic economic principles, economic models often need to be tightly specified in such a way that we should *not* expect model errors to be unsystematic *iid*. Such models will be statistically rejected at a given significance level if the test is sufficiently powerful. Economists, therefore, to an increasing extent analyse and evaluate economic models empirically using methods that are better suited for misspecified models than statistical hypothesis tests. Instead of testing whether the model is statistically rejected at a given significance level, these methods measure the degree of misspecification (e.g. the *magnitude* of pricing errors in asset pricing models) and they analyse in which dimensions - and to what extent - the model fits the data, and in which dimensions it does not.

Despite McCloskey & Ziliak's strong focus on 'real errors', and although they explicitly refer to the fact that models are neither 'true' nor 'false' (Ziliak and McCloskey, 2008, p.52), they are apparently not aware of those fields in economics where models are *not* treated as either 'true' or 'false', and where the distinction between statistical and economic significance is very clearly spelled out - and has been increasingly so over the last more than 20 years. I mention three such disciplines: sto-

chastic general equilibrium models; linear rational expectations models; and asset pricing models. In all three areas there is a clear recognition that statistical (in)significance does not necessarily imply economic (in)significance, and that standard statistical measures of fit may not be the most informative way of evaluating a given model. I provide examples of this from the literature; papers and textbooks that have had a large impact on how empirical researchers in these areas think about modelling. Thus, I provide direct evidence against one of McCloskey & Ziliak's central claims, namely that almost all economists - even today - confuse statistical and economic significance.

Besides this main point, I have a number of additional comments to some of McCloskey & Ziliak's claims and statements. Some of the comments are in line with the comments made by Hoover & Siegler. In particular, I argue that conventional statistical testing can be used as an effective and valuable tool in obtaining well-specified and parsimonious statistical models that subsequently can be used as input in the quantification of economic models. However, here it is important to distinguish between the statistical model and the economic model; treating the latter as a 'null model' to be tested statistically is not particularly informative. I also criticise McCloskey & Ziliak for their insistence on forgetting standard errors and focusing on parameter estimates, while they provide almost no discussion of how to obtain good reliable parameter estimates. Finally, I provide a specific example of a research area in economics - financial asset return predictability - where the distinction between statistical and economic significance is well appreciated, but which also shows how statistical tests have contributed to our substantive economic understanding. Findings of *statistically* significant return predictability set the stage for a deeper inquiry into the nature of return predictability and its implications for e.g. portfolio choice and asset pricing, and thereby contributed in changing our minds about the functioning of financial markets.

Before getting to the main body of the paper, I would like to state from the outset that I fully agree with McCloskey & Ziliak's point that statistical (in)signi-ficance does not necessarily imply economic (in)significance, and that good empirical research in economics should discuss economic significance one way or the other (on the other hand, I agree with Hoover & Siegler that today this is in fact what most economists do, see further below). I also agree with McCloskey & Ziliak that real scientific progress in economics - how we change opinion on how the economy works - is achieved mainly through common sense, elegant theories, historical perspective, and long and disciplined *conversations* among scholars, i.e. how *pursuasive* we are in our discussions

(the 'rhetoric' of economics), c.f. McCloskey (1983) and McCloskey and Ziliak (1996).[1] There is no 'objective' method or standard (like the 5% significance level) that *in itself* can decide for us. And no change from one paradigm to another has been driven mainly by 'statistical significance at a 5% level'. But I believe that statistical hypothesis testing can be used as one of several important tools, and I think there are examples in economics where statistical hypothesis testing has played an important role in moving from one paradigm to another (see below).

2 Scientific evaluation of misspecified models

At one important point it seems that Hoover & Siegler and McCloskey & Ziliak talk at cross-purposes. Hoover & Siegler confine almost exclusive attention to sampling or statistical error and error that can be eliminated through specification testing, while McCloskey & Ziliak emphasize 'real error', error that cannot be eliminated through specification testing.[2,3] Hoover & Siegler work under the premise that there is a 'true' specification that can be recovered by specification testing, and they advocate the socalled 'LSE approach' originating from the London School of Economics (Hoover and Siegler, 2008a, p.26). Naturally, Hoover & Siegler do not search for the literally true data-generating-mechanism; they are aware that models are only approximations to reality. They search for models in which the errors are basically white noise. However, many economists are sceptical towards this approach, see e.g. Faust and Whiteman (1997). The LSE methodology obtains empirical models with well-behaved *iid* error terms, but those models are often so complex

[1] I do, however, find McCloskey & Ziliak's rhetoric sometimes a bit tiring, for example when stating that "all the econometric findings since the 1930s need to be done over again." (McCloskey and Ziliak, 2008, p.47), or: "If null-hypothesis significance testing is as idiotic as we and its other critics have so long believed, how on earth has it survived?" (Ziliak and McCloskey, 2008, p.240).

[2] Hoover & Siegler refer to 'specification tests' as tests for no structural breaks and model errors being serially uncorrelated, homoscedastic, normally distributed, and at all white noise (c.f. Hoover and Siegler, 2008, p.24). In econometric textbooks such tests are often referred to as 'misspecification tests' while 'specification tests' denote tests within the econometric model based on the assumption of correct specification (see e.g. Spanos, 1986, section 19.5). Thus, *given* white noise model errors (which is tested using 'misspecification tests'), 'specification tests' are used to test hypotheses on e.g. regression coefficients. In the following I will follow Hoover & Siegler and instead use 'specification tests' to denote tests that are used to secure unsystematic *iid* model errors.

[3] Although McCloskey & Ziliak refer to 'real error' and its importance several times in their writings, they don't provide a precise definition of what they mean by it in relation to economic modelling. But from Ziliak and McCloskey (2008, pp.7 and 24) I infer that by 'real error' they refer to non-sampling error that leads to models being inherently misspecified.

with many variables (including dummy variables) and complicated lag-structures, and with only loose connection to economic theory, that many find them difficult to interpret economically in a consistent manner. Just like estimation and measurement of unknown quantities based on data samples need to be based on relevant statistical theory, *economic theory* is the vehicle through which economists make consistent interpretations of the very complex dynamics and interactions we observe among almost all economic variables. Empirical economic modelling without close ties to established economic theory is like estimation without close ties to statistical estimation theory. This is not to say that the only purpose of empirical modelling is to confirm economic theory. Rather, the purpose is to investigate the empirical content of economic theory: how well does the theory explain the data in this and that dimension? And then one can look at those dimensions where the empirical performance is bad, and that evidence can be used to reformulate the theory in order to perform better. Hence, the aim of the empirical modelling exercise is not to confirm the validity of a theory, but to investigate in which dimensions the theory needs to be modified, and always remembering that we will never discover the 'true' relation.

As an alternative to the LSE approach, a modelling framework based on economic theory and with explicit acknowledgement of inherent model misspecification (i.e. non-negligible systematic deviations between model and data), has been adopted by many empirical researchers in macroeconomics and finance. McCloskey & Ziliak are very critical towards practice in macro and finance (Ziliak and McCloskey, 2008, pp.69-70, 76). But, in fact, in these fields it is explicitly acknowledged that economic models are inherently misspecified, i.e. contain 'real errors' in the language of McCloskey & Ziliak. Research on *rational expectations* models is an example. McCloskey & Ziliak rebuff the whole area of rational expectations macroeconometrics as uninteresting and with "no scientific findings", with reference to the work of Lucas, Sargent, and others, from the 1970s and beginning of the 1980s (Ziliak and McCloskey, 2008, p.108). This is peculiar because since the 1980s the research *methodology* within the rational expectations paradigm has progressed in a way that seems to fit exactly with McCloskey & Ziliak's prescriptions. Let me mention two very influential sub-fields within the area of macroeconomics where rational expectations have played a central role: Dynamic Stochastic General Equilibrium (DSGE) models - previously denoted Real Business Cycle (RBC) models -, and Linear Rational Expectations (LRE) models. In both of these areas it is explicitly acknowledged that models are by construction misspecified and, hence, that empirical evaluation by statistical significance tests (e.g. 'goodness-of-fit' tests of

overidentifying restrictions and tests for 'correct' specification) become less informative. The models don't pretend to be 'true' so they are expected to be rejected by statistical tests if these tests are sufficiently powerful.

In the DSGE/RBC literature these concerns were expressed from the very beginning. Kydland and Prescott write in their highly influential paper from 1982: "We choose not to test our model ... this most likely would have resulted in the model being rejected, given the measurement problems and the abstract nature of the model." (Kydland and Prescott, 1982, p.1360). Instead, calibration and simulation techniques are used in computational experiments where model-generated moments of key variables are compared to moments of actual variables, and where focus always is on *quantitative* questions - questions of "how big something is", c.f. Kydland and Prescott (1996, p.75). This is exactly what McCloskey & Ziliak call for! A typical *quantitative* research question in this area is: "How much would the U.S. postwar economy have fluctuated if technology shocks had been the only source of fluctuations?" (Kydland and Prescott, 1996, p.77). Another very influential contribution in this area is Mehra and Prescott's (1985) 'equity premium puzzle' paper which shows an *economically* significant puzzle (the high equity premium without high equity risk) without using statistical tests.[4] Thus, McCloskey & Ziliak's claim that rational expectations macroeconomics haven't produced any scientic findings is obviously not true! The model evaluations involved in the early calibration and simulation exercises from the 1980s and 1990s have been heavily criticised for being ad hoc and with no firm statistical foundation. However, recent DSGE research has extended and elaborated on these basic calibration and simulation exercises by explicitly introducing *loss functions* and by using Bayesian econometric procedures for parameter estimation, model evaluation, and model comparison, while continuing to consider the underlying economic model as inherently misspecified, see e.g. Schorfheide (2000) and Fernández-Villaverde and Rubio-Ramírez (2004) (see also

[4]In the equity premium puzzle literature it is a common observation that the standard consumption-based asset pricing model with time-separable power utility cannot explain the equity premium puzzle without running into a 'risk-free rate puzzle', and that the model cannot account for the time-varying counter-cyclical nature of expected returns. This has lead to the development of alternative models, for example the habit persistence model of Campbell and Cochrane (1999). Both models are typically rejected statistically, but the Campbell-Cochrane model performs better because it accounts for more of the empirical facts we observe. And note that Campbell and Cochrane themselves do not evaluate their model using statistical tests. They do a calibration/simulation exercise together with traditional estimation of parameters, acknowledging that their model is inherently misspecified. I return to asset pricing models and the equity premium puzzle later in this section.

Watson, 1993, and Diebold et al., 1998, for some early attempts to develop rigorous standards for relating calibrated models to data). Again, exactly in accordance with what McCloskey & Ziliak consider good scientific practice. I emphasize that whether one agrees or disagrees with the DSGE research program is not the issue here. There may be reasons for McCloskey & Ziliak not liking DSGE modelling, but adherence in this modelling framework to statistical significance cannot be one of them!

A related research field is the use of simple linear rational expectations (LRE) models to explain key macroeconomic and financial variables like consumption, labour demand, money demand, inventories, the balance of payments, stock prices and interest rates. LRE models have been very popular since Hansen and Sargent's work in the 1970s and early 1980s, see e.g. the very influential paper by Hansen and Sargent (1980). In the early work in this area (including Hansen and Sargent's) formal statistical tests of model-implied overidentifying restrictions played an important role. A key element in the empirical evaluation of LRE models was the testing of a null hypothesis that the model is 'true' in the sense that the discrepancy between model and data is due to only sampling error. This is probably the reason for Ziliak and McCloskey's (2008, p.108) complete rejection of all of rational expectations macroeconomics. But apparently they have stopped reading the literature with the qouted book and papers by Lucas and Sargent from 1981.[5] Soon after it was realized in this field that taking a LRE model to be the null hypothesis and rejecting or accepting it based on a computed p-value (i.e. treating the model as either 'true' or 'false'), is not very informative. Probably the most influential paper that directly expresses this concern is Campbell and Shiller (1987, p.1063): "... a statistical rejection of the model ... may not have much *economic* significance. It is entirely possible that the model explains most of the variation in y_t even if it is rejected at a 5% level." (Italics added). Campbell and Shiller propose an alternative metric for empirical evaluation of a LRE model. The idea is to estimate a Vector-AutoRegression (VAR) for the variables in the model and from the VAR parameter estimates to generate a model-implied time-series for the 'endogenous' variable, y_t, which can then be compared with the actual time-series for the variable y_t. A graphical

[5]In fact, Lucas never really advanced the Hansen and Sargent (1980) approach, on the contrary, see e.g. Lucas (1987). Lucas' views on statistical testing of economic models are basically identical to the views expressed by Kydland and Prescott (1982). According to Sargent, Lucas already in the early days of rational expectations macroeconomics expressed the concern that "likelihood ratio tests are rejecting too many good models", c.f. Evans and Honkapohja's (2005) interview with Sargent.

time-series plot of actual and 'theoretical' y_t is an important part of the comparison, and the methodology can be used to obtain an explicit measure of the *magnitude* of deviations from the underlying LRE model, i.e. *how much* (measured as a percentage) of the variablity in y_t is explained by the model and how much is due to model noise (see Engsted, 2002, which also lists some of the many published applications of Campbell and Shiller's methodology). Thus, this methodology explicitly addresses the problem of 'real error' as opposed to only sampling error; however, McCloskey & Ziliak pay no attention to it.

Yet another area where the limitations of statistical hypothesis testing and the distinction between statistical and economic significance is clearly acknowledged is *asset pricing* in finance. Of course, this area is closely connected to the rational expectations paradigm but it has grown into a large and individual sub-field of economics. Through the 1970s and 1980s empirical evaluation of asset pricing models in most cases followed the traditional statistical approach of testing overidentifying restrictions implied by the models. A very influential example of this kind of research is Hansen and Singleton's (1982) *Generalized Method of Moments* (GMM) approach for testing the consumption-based capital asset pricing model (C-CAPM). However, during the 1990s and 2000s focus shifted towards an approach with more emphasis on assessing model performance and measuring the magnitude of pricing errors. Hansen and Jagannathan (1991, 1997) are two very influential papers in this progression, and Cochrane and Hansen (1992, p.122) summarize the main point very clearly: "Statistical measures of fit such as a chi-square test statistic may not provide the most useful guide to the modifications that will reduce pricing or other specification errors ... Also, application of the minimum chi-square approach to estimation and inference sometimes focuses too much attention on whether a model is perfectly specified and not enough attention on assessing model performance." Hansen and Jagannathan (1991) propose a graphical method for evaluating a given asset pricing model. In contrast to the traditional statistical approach, the method provides information on the dimensions in which the model fails and it points in direction to which the model needs to be modified. In their 1997 paper Hansen and Jagannathan propose an explicit measure of the *magnitude* of pricing errors of a given inherently misspecified asset pricing model. Both papers have been highly influential in modern empirical research in finance, but they (and their many followers) are completely neglected by McCloskey & Ziliak.[6]

[6]In addition to the work with Jagannathan, Hansen has - together with Sargent - for more than 10 years worked on 'robust control' in which decision makers take into account model misspecification, see e.g. Hansen and Sargent (2007). Thus, both

The explicit acknowledgement of the limitations of statistical hypothesis testing and the important distinction between statistical and economic significance have also found their way into modern textbooks, in contrast to what McCloskey & Ziliak claim. Let me mention one example of an empirical and econometrically oriented graduate textbook: Cochrane's *Asset Pricing* from 2001. This book has already become a standard reference for students and empirical researchers in finance. Here are some examples from the book:

On page xvi in the Preface, Cochrane emphasizes that empirical methods in the end "... evaluate the model by examining *how big* [the] pricing errors are." (Italics added).

On pages 194-196 (chapter 10), Cochrane is very careful in stating that the socalled J_T test measures pricing errors in a *statistical* sense: "The J_T test asks whether [pricing errors] are "big" by *statistical* standards" (p.196). (Italics added).

On pages 210-219 (chapter 11), Cochrane is very careful in distinguishing between *statistical* and *economic* measures of fit (see in particular pp.210, 215, 218).

On pages 291-305 (chapter 16), Cochrane provides a detailed discussion of the limitations of using statistical significance tests in evaluating inherently misspecified models (see especially pp.303-305, where in fact there are explicit references to McCloskey! Cochrane makes the interesting observation that "McCloskey's ideas are not popular in the finance and economics profession. Precisely, they are not popular in how people *talk about* their work, though they describe well how people *actually do* their work." (p.304)).

On pages 434-441 (chapter 20), Cochrane gives a nice graphical (no significance tests!) illustration of the failure of one kind of asset pricing model and how an alternative model works better.

On pages 455-485 (chapter 21), Cochrane illustrates the equity premium puzzle and the 'fit' of various asset pricing models, almost without using statistical significance tests.

In Cochrane (2006, p.17), Cochrane refers to the paper by Fama and French (1996) as one that "... for better or worse, defined the methodology for evaluating asset pricing models for the last 10 years. ... where in the 1980s papers would focus entirely on the probability value of some overall statistic, Fama and French rightly got people to focus on the spread in average returns, the spread in betas, and the *economic size of the pricing errors*. Remarkably, this, the most successful model since the CAPM, is decisively *rejected* by formal tests. Fama and French

Hansen and Sargent explicitly acknowledge the problems with their initial rational expectations models of the 1970s and early 1980s.

taught us to *pay attention to more important things than test statistics.*" (Italics added). Inspection of various issues of the Journal of Finance since the mid 1990s confirm Cochrane's statement.

As a final example of empirical research where statistical hypothesis testing plays only a minor role, let me mention *optimal asset allocation*. There is a voluminous literature, starting in the 1990s, on optimal portfolio choice for long-term investors. Here, researchers estimate parameters for the dynamic evolution in returns and their state variables, insert them - together with calibrated utility parameters - into optimal dynamic portfolio equations, and measure the economic significance of various portfolio choices by computing utility losses of excluding this and that asset. Statistical significance tests play only a very minor role. And note that here there is an explicit loss function! Recent examples of this kind of research is the book by Campbell and Viceira (2002), and the paper by Campbell, Chan, and Viceira (2003), both already widely cited.

It should be clear from the above examples that McCloskey & Ziliak are not correct when they claim that almost no empirical researchers in economics are able to distinguish properly between statistical and economic significance.[7] This is not to say that economists have followed McCloskey & Ziliak in totally dismissing statistical hypothesis testing. Most (including myself) continue to consider regression standard errors and statistical tests as useful tools in empirical modelling, but we clearly acknowledge the limitations of such tools and we supplement the statistical tools with economically more informative assessments. In the next section I discuss in more detail McCloskey & Ziliak's arguments for completely dismissing statistical hypothesis testing.

[7] Underlying McCloskey & Ziliak's claim is a survey analysis of practice in empirical papers published in the American Economic Review during the 1980s and 1990s (see McCloskey and Ziliak, 1996, and Ziliak and McCloskey, 2004). Hoover and Siegler (2008a) criticise McCloskey & Ziliak for erroneously omitting a significant number of relevant articles from the American Economic Review in their two 'questionnaire' surveys, and for inaccurate readings and tendentious interpretations of the articles included. In addition, McCloskey & Ziliak's surveys can be criticised for containing only papers that use statistical tests in connection with regression analysis. Thereby they potentially omit many relevant papers. As I have argued above, researchers in macroeconomics often don't use statistical significance tests because they are highly critical of such tests; instead, they use calibration and simulation to evaluate their models. And in finance - where researchers also clearly understand the difference between statistical and economic significance, as I have argued -, econometric methods other than simple regression analysis are often used. Such papers are apparently not included in McCloskey & Ziliak's surveys. Thus, most probably McCloskey & Ziliak erroneously leave out papers that are doing exactly what they call for.

3 Is statistical hypothesis testing useless? And what about the properties of parameter estimates?

In much of McCloskey & Ziliak's writings they focus on the case where a statistical hypothesis test is used to test a null hypothesis of no effect, i.e. a parameter value equal to 0. Basically they find such a test useless as a *scientific* tool. They state that scientists should not be interested in whether there is an effect, but how big it is: "Existence, the question of whether, is interesting. But it is not scientific" (Ziliak and McCloskey, 2008, p.5). The question of whether is a *philosophical* question, they say. Instead, "Statistics, magnitudes, coefficients are essential scientific tools" (Ziliak and McCloskey, 2008, p.1). In addition, they advocate Bayesian procedures in econometric modelling and the Neyman-Pearson framework in which explicit loss functions play an important role. Rejecting or non-rejecting a null hypothesis should not be based on some arbitrary and conventional significance level, but on the relative costs and benefits (measured in *economically* relevant terms) of either rejecting or not rejecting. They take this line of reasoning to argue that also standard errors (or precision measures in general) of parameters should be judged in terms of explicit loss functions: "The sheer probability statement about one or two standard errors is useless, unless you have judged *by what scale a number is large or small for the scientific or policy or personal purpose you have in mind.* This applies to the so-called 'precision' or 'accuracy' of the estimate, too" (McCloskey and Ziliak, 2008b, pp.43-44). On the whole McCloskey & Ziliak put strong focus on the magnitude (what they call 'oomph') of estimated coefficients but don't seem to bother much about the statistical uncertainty surrounding such estimates: "Oomph, Not Precision, Selects the Best Model" and "Precision usually does not pick the right dimension for comparison. Oomph does." (Ziliak and McCloskey, 2008, pp.48-49).

McCloskey & Ziliak's focus is quite narrow in that they typically discuss hypothesis testing in the context of a single parameter, and they assume that this single parameter has an important economic meaning. In this narrow context it is obviously true that one should not include or exclude the parameter based alone on whether it is statistically significant at an x% level. But there are many examples in economics where the magnitude (or even sign) of an estimated coefficient is not particularly interesting in itself. Vector-AutoRegressions (VAR's), for example, are often used as 'reduced form' summaries of the dynamic evolution over time in a number of variables, from which e.g. impulse responses or long-term relations or optimal portfolio choices are inferred, which are then interpreted economically. But the magnitude and sign of the

individual VAR coefficient are not particularly interesting. Specification tests are used here to address the statistical adequacy of the VAR. Of course, in such tests the '5% rule' (or 1% or 10% for that matter, depending on the sample size and the trade-off between size and power) is just a *convention*, but as Hoover & Siegler argue this does not necessarily make it useless or ineffective. For example, such a rule can be useful in reducing a large-scale model, e.g. a reduced-form model where the individuel parameters are not particularly interesting *per se*, into a more parsimonious model. As Hoover and Siegler (2008b, p.58) point out, conventions are often employed in *scientific judgement*. To take a specific example, in analyzing the relationship between stock prices and expected future dividends, the stock price is measured at a point in time, t, while dividends are paid sometime within the period t. Thus, the researcher needs to make the assumption that time t dividends are known to the market at the precise time point that prices are measured (c.f. Campbell and Shiller, 1987, p.1074). An often used convention in such studies is to measure prices at either the beginning or the end of year t and then let dividends (which in theory should be measured contemporaneously with prices) be included with a one-year lag (if prices are measured at the start of the year) or without a lag (if prices are measured at the end of the year). In the same vein, a standard timing convention in consumption-based asset pricing models, for example, is to assume - completely arbitrary - that observed consumption takes place at the *beginning* of the period. Scientific research cannot be done without sometimes referring to conventions. Naturally, significance tests should not substitute for proper robustness check of the *economic* implications of changes in the econometric model. Do the impulse responses or long-run relations or portfolio allocations change fundamentally by changing the variables and the lag-length in the VAR or by changing the sample periods? But such robustness checking is in fact an integral part of most applied econometric research.

The above defense of specification testing does not contradict the arguments I made in section 2 for putting *less* weight on statistical testing of inherently misspecified models. It is important here to distinguish between the kind of model one is working with. For example, in the Campbell-Shiller methodology for evaluating linear rational expectations models, described in section 2, there are two kinds of models. One is the *economic* rational expectations model in which one variable, y_t, is determined af the present discounted value of expected future values of another variable (or set of variables), x_t. And the other is the *econometric* model, a VAR model for y_t and x_t that summarizes the dynamic evolution of - and interaction between - y_t and x_t, and which is used

to generate expected future values and the 'theoretical' (model-implied) time-series for y_t. In this setting, because we acknowledge that the economic model is only a crude approximation to the 'true' mechanism that generates y_t, a statistical test of whether actual y_t deviates significantly from theoretical y_t has only limited informative value. However, statistical specification testing of the underlying econometric VAR model is more informative because the validity of the constructed theoretical y_t depends crucially on the VAR system capturing the essential dynamics over time and interactions between y_t and x_t.[8]

The '5% rule' is of course meaningless if used mechanically and thoughtlessly in each and every application. However, as argued above, in *some* cases such a rule can be applied as a useful convention. In many areas of empirical research practice has shifted from designating stars to numbers significant at a particular level to instead reporting p-values. Denote by β and $\widehat{\beta}$ the unknown true value of a parameter and the sample estimate of it, respectively. A p-value for a null hypothesis of $\beta = 0$ shows the probability of observing a value equal to $\widehat{\beta} > 0$ (or larger) if, in fact, $\beta = 0$, given the sample. If the p-value is low it means that there is a low probability of obtaining $\widehat{\beta}$ if the true value of β is 0. Thus, low p-values usually lead to rejection of $\beta = 0$. The advantage of reporting p-values instead of merely stating significance at a given significance level is that the reader can compare it to his or her own subjective critical significance level. It is not entirely clear what McCloskey & Ziliak think of p-values. On the one hand they dismiss p-value calculation: "The scientifically relevant question is a question of how big the parameter of interest is, not the Fisherian question of how probable the data are, given the null hypothesis, a purely sampling problem" (Ziliak and McCloskey, 2008, p.95). On the other hand, on page 99 they seem to accept such p-value calculation: "A fair question to ask ... is *how noisy*? Just how weak was the signal-to-noise ratio, assuming that one thinks the measure is captured by the calculation of sampling error? The answer underscores the arbitrariness of Fisher's 5 percent ideology - the Type I error was about 12 percent ($p \leq .12$). That is to say, the 4.29 benefit-cost ratio was ... statistically significant at about the .12 level. In other words, the estimate was not all that noisy. A pretty strong signal ...". Thus, given that we accept using sampling error (i.e. standard error) to measure signal-to-noise, McCloskey & Ziliak

[8]Note that the Campbell-Shiller methodology does *not* imply that high R^2 and strongly statistically significant parameters in the VAR model lead to close comovement of actual and theoretical y_t. It depends on the signs and relative magnitudes of the parameters. Thus, high fit in the VAR model does not necessarily imply high fit in the economic model.

find it relevant and informative to compute p-values. The essential element in the above quotation is "assuming that one thinks the measure [signal-to-noise ratio] is captured by the calculation of sampling error". So, the whole point can be boiled down to whether sampling error is an adequate measure of noise. In the presence of inherent model misspecification ('real error') sampling error is obviously not an adequate measure of noise, but is is noteworthy that here McCloskey & Ziliak in fact leave the door half open for using standard errors and significance testing.

McCloskey & Ziliak's insisting on evaluating each and every finding in the context of an explicit loss function ("Fisherian significance without a loss function is ordinarily useless for science", McCloskey and Ziliak, 2008, p.46) is in theory appealing but in practice often inapplicable. In principle they are right in saying that behind every rejection/non-rejection at a given significance level, and behind most (but not all, see footnote 9 below) measures of precision in terms of a standard error, lies an implicit loss function that depends on the underlying economic purpose of the analysis. But often a parameter estimate has many different uses and, hence, many different relevant loss functions. A research paper cannot possibly list them all. What's wrong with reporting the parameter estimate and it's standard error, and then let it be up to the reader to apply these in his or her specific context or loss function? As Hoover and Siegler (2008b, p.59) put it: "Whose loss function should Newton have consulted?".

McCloskey & Ziliak's repeated suggestion of focusing on the size of parameter estimates while forgetting the standard error ("Oomph, Not Precision, Selects the Best Model") leaves the question of how to obtain good reliable parameter estimates. When can we rely on a parameter estimate? It almost seems like McCloskey & Ziliak are ready to trust *any* estimate. In any case, they provide almost no discussion of what - in their opinion - constitute good reliable estimates. In McCloskey and Ziliak (2008, p.48) they write about deciding about the importance of a variable by use of statistical hypothesis testing and say that: "It's the wrong way to decide, and leaves the wrong variables in the regressions, and results in *biased and inconsistent* estimates of the coefficients." (Italics added). And In Ziliak and McCloskey (2008, p.122) they write that "continuing to decorate our articles with stars and t's and standard errors while failing to interpret size - is to discard our best *unbiased* estimators ...". (Italics added). And that's it! No further discussion. From the two qoutes one can infer that McCloskey & Ziliak consider *unbiasedness* and *consistency* to be properties to strive for in parameter estimation. I take it that they mean unbiasedness and consistency in the traditional statistical meaning, i.e. the mean of the estimate should equal the true value

for unbiasedness and the probability limit should equal the true value for consistency. But then they must acknowledge the need to secure that the assumptions for unbiasedness and consistency hold true. How do we do that without using statistical tests? For example, in the presence of non-stationary variables a necessary condition for consistency of OLS parameter estimates is that the variables are *cointegrated* in the sense of Engle and Granger (1987). But Ziliak and McCloskey (2008, p.111) dismiss tests of cointegration. Thus, apparently they consider the 'spurious regression' problem (Granger and Newbold, 1974, and Phillips, 1986) not to be a problem at all. So, no matter what the time-series properties of the variables, they will always trust OLS parameter estimates? Or, do they mean that cointegration can be inferred without using some kind of statistical test? In any case: how do we secure that our parameter estimates are well-defined and have 'good' properties? Hoover and Siegler (2008a) provide further discussion of this issue, and they also criticise McCloskey & Ziliak for sweeping under the carpet potential problems associated with parameter measurement and estimation.[9] As noted in the previous section, McCloskey & Ziliak emphasize many times that 'real error' is more important than pure sampling error. But this just reinforces the need to discuss the properties of parameter estimates. How reliable are regression estimates in the face of 'real error'? McCloskey & Ziliak are completely silent about this.

4 A case study: Return predictability

In this section I provide a specific example of a research area in economics in which the arguments of McCloskey & Ziliak can be made concrete, and where the distinction between statistical and economic significance is well appreciated, but which also shows how statistical tests have contributed to our substantive economic understanding. Return predictability is briefly mentioned by McCloskey & Ziliak as an area where the distinction between statistical and economic significance is particularly concrete. As Ziliak and McCloskey (2008, p.112) note, statistically *in*significant return predictability may make you rich, and statistically significant predictability may not be large enough to cover transaction costs.

[9] Hoover and Siegler (2008a, pp.17 and 24) provide an explicit example showing the danger of relying on poorly measured parameters. The example shows that "eliminating a variable from a regression may be a reasonable thing, despite its *economic* significance." (p.24). (Italics added). The example illustrates how strong multicollinearity among regressors, which destroys parameter estimates, can be diagnosed using the standard errors of the estimates, in a way that is independent of the underlying 'loss function', in contrast to what McCloskey & Ziliak claim.

There is by now a voluminous literature on asset return predictability. The old paradigm of finance stated that asset returns are unpredictable and that capital markets are informationally efficient. This view was prevalent up to around the mid 1980s. But a plethora of empirical research since then has documented that returns contain statistically as well as economically significant predictable components and, interestingly, theorists have shown that such predictability does not necessarily imply that markets are inefficient. It has long (at least since the 1980s) been recognised that the *Efficient Markets Hypothesis* cannot be tested since it involves a double-hypothesis: informational efficiency *and* correctness of the underlying equilibrium model generating expected returns. Thus, in this field there is not much of the 'falsificationism' or 'Popperian philosophy' or 'logical positivism' that McCloskey & Ziliak so strongly dislike, c.f. Ziliak and McCloskey (2008, pp.149-153). But statistical testing for predictability has played an important role in reaching the new paradigm of return predictability. And not just mindless 't-tests' popping out of standard econometrics software programs, but carefully conducted tests (including Monte Carlo and bootstrap analyses) taking into account non-normality, asymmetric distributions, small-sample bias, etc. It's interesting to look at how research in this area has progressed over time. The first studies in the 1980s focused mainly on documenting statistically significant predictability using asymptotic tests (e.g. Fama and French, 1988). Then during the 1990s these findings were challenged by research (e.g. Nelson and Kim, 1993, Stambaugh, 1999) showing that accounting for small-sample bias leads to statistically *in*significant predictability. Finally, over the last 10 years focus has shifted to examining the *economic* significance of statistically large or small predictability, see e.g. Xu (2004) and Cochrane (2008). But note that it all started with research where statistical significance was the main focus. These studies set the stage for a development that in the end changed our minds regarding asset return predictability. Thus, a finding which is 'statistically significant at a 5% level' does not by itself change our mind, but it can initiate subsequent research that ultimately does.

The literature on return predictability provides a lot of what McCloskey & Ziliak call for. They call for more extensive use of *simulations* "to determine whether the coefficients are reasonable" (Question 17 in their survey of the American Economic Review), and they want researchers to report (or at least discuss) the *power* of the tests they use (Questions 8 and 9 in the survey). Ziliak and McCloskey (2008, p.60) describe the virtues of the 'β-scientist' (as opposed to the 'α-scientist' who is only concerned with statistical significance), one who is "concerned with empirical interpretation and judgement .... small-sample

experience of life.... non-normal distributions ... Type-II error - the power of tests -". But, in fact, all this is there, in todays empirical work. Simulations and power considerations occupy much of empirical financial economics. In the return predictability literature it has become standard to use Monte Carlo and bootstrap simulation to take into account non-normal and asymmetric distributions, and to account for small-sample bias of parameter estimates, see e.g. Nelson and Kim (1993) which is an early standard reference in this field. Power properties of tests are usually discussed, if not analysed. There is a whole section (15.2) in Cochrane's (2001) textbook on simulation and power calculations in empirical finance.[10]

Research on return predictability has documented an interesting difference between short and long run effects, and how small and difficult-to-measure short-run predictability implies large and better-measured long-run predictability. The nature of asset returns is such that small short-horizon predictability build up to large long-horizon predictability, and short-horizon predictability is a necessary condition for long-horizon predictability (c.f. Cochrane, 2001, ch.20). We are not particularly interested in the magnitude of short-horizon predictability because it is almost impossible to benefit from. Rather, we are interested in whether there is an effect, because even very small short-run predictability builds up to large long-run predictability. Therefore, in e.g. portfolio allocation applications where regression models are used to characterize short-horizon predictability, it has become standard practice to keep in the model predictors with 'small' coefficients that are statistically *in*significant at conventional significance levels, because even small short-horizon coefficients have large effects on long-term optimal portfolio choice. Here, researchers are well aware that a statistically *in*significant short-run effect may be economically significant in the sense that it has important long-run effetcs. The portfolio allocation example also serves to illustrate the distinction between the econometric model and the economic model, c.f. section 3. The *econometric* model captures the basic return predictability properties of the data, and this model needs to be econometrically well-specified which can be achieved through specification testing. The *economic* model - the specific portfolio choice model based on constrained utility maximization - on the other hand, uses as input in the quantification of the model the parameter estimates from the econometric model. But the economic model

[10] Another example is the large field of unit root and cointegration testing, which are standard tools in empirical macroeconomics - and completely dismissed by McCloskey and Ziliak. Here, many studies have investigated the power properties of such tests (see e.g. the well-known study by Elliott, Rothenberg, and Stock, 1996).

is highly stylized and based on many simplifying assumptions (it has to, in order to be economically interpretable), thus it is inherently misspecified and, hence, testing whether this model is 'true' with a statistical hypothesis test is not particularly useful. The *statistical/econometric* model is judged by *statistical/econometric* criteria, while the *economic* model is judged by *economic* criteria. Of course, if statistical methods are used in estimating the economic model parameters, then the statistical assumptions for validity of the estimates should be fulfilled. But this is also why in this field the preference parameters in the economic model are typically not estimated but calibrated in the same way as in the DSGE/RBC literature referred to in section 2.

In the finance literature, *economic* considerations have even been used to sharpen the *statistical* evidence of return predictability. For example, Lewellen (2004) uses the stylized fact of strong persistence in dividend yields (a standard return predictor) together with a *prior* of no speculative bubbles in stock markets to obtain more precise estimates of short-run stock return predictability. Cochrane (2008) - in the same vein - adds lack of dividend predictability (which is another stylized fact in finance - statistically as well as economically) to the prior information of dividend yield persistence and no bubbles, to document strongly statistically significant return predictability.[11] Thus, in this literature there is a clear tendency of using *economic* arguments to guide formulating relevant *statistical* measures and tests of predictability. Without being Bayesian, there is a clear Bayesian flavour to these analyses.

Thus, the literature on return predictability illustrates the importance of distinguishing between statistical and economic significance, but it also shows that McCloskey & Ziliak are wrong in stating that almost no empirical researchers in economics understand this. Todays empirical research in finance is conducted mainly by β-scientists! And in this area findings of *statistically* significant return predictability in fact set the stage for a deeper inquiry into the nature of return predictability and its implications for e.g. portfolio choice and asset pricing, and thereby contributed in changing our minds about the functioning of financial markets.

5 Concluding remarks

McCloskey & Ziliak believe that past and current practice in applied econometrics is so fundamentally flawed in almost all cases that "all the econometric findings since the 1930s need to be done over again."

[11] Cochrane (2008) from the outset distinguishes between statistical and economic significance. The very first section in the Introduction of his paper is labelled "*Economic Significance*". The second section is labelled "*Statistical Significance*".

(McCloskey and Ziliak, 2008, p.47). They base this statement on the claim that almost no economists that apply statistical and econometric methods in their research are able to distinguish between statistical and economic significance. In this paper I have provided several examples from the economics disciplines showing that this claim is false. I have mentioned three specific fields from macroeconomics and finance - areas that McCloskey & Ziliak explicitly refer to as areas where the 'significance mistake' is particularly pronounced - where researchers are, and have been for several years, aware of the simple fact that statistical significance is neither necessary nor sufficient for economic significance: Dynamic stochastic general equilibrium models, linear rational expectations models, and asset pricing models. In these areas (and I'm sure in other areas as well) it is clearly acknowledged that economic models are inherently misspecified and, hence, that treating such models as a 'null hypothesis' to be tested statistically is not very informative. Thus, in these areas researchers pay close attention to what McCloskey & Ziliak call 'real error', but surprisingly this evidence is completely neglected by McCloskey & Ziliak.

Furthermore, I have argued that putting little weight to statistical testing of restrictions implied by the *economic* model structure, does not imply that statistical tests are without value in economic scientific research. On the contrary. *Statistical* (or *econometric*) models often serve as input into the quantification of economic models, and here it is important that the statistical model is well-specified. Specification tests are a useful tool in obtaining that.

I have also argued that although McCloskey & Ziliak are right in saying that no statistical test *in itself* has changed our mind about a substantive matter, such tests occasionally spur research that does. The last 20-25 years research on asset return predictability is an example where initial findings of statistically significant return predictability set the stage for subsequent research that ultimately changed the minds of most financial economists about return predictability and its implications for asset pricing and asset allocation.

Finally, I have criticised McCloskey & Ziliak for urging us to focus on the size of parameter estimates and forgetting standard errors, while they provide almost no discussion on how to obtain good reliable parameter estimates. They seem to be perfectly happy with standard regression estimates. This is surprising in light of their strong focus on 'real error' as opposed to pure sampling error. Such 'real error' will in many cases invalidate standard regression estimates. Surprisingly, McCloskey & Ziliak are silent about this problem.

6 References

Campbell, J.Y., and Cochrane, J.H. (1999), "By force of habit: A consumption-based explanation of aggregate stock market behavior", Journal of Political Economy 107, 205-251.

Campbell, J.Y., and Shiller, R.J. (1987), "Cointegration and test of present value models", Journal of Political Economy 95, 1062-1088.

Campbell, J.Y., and Viceira, L.M. (2002), "Strategic Asset Allocation: Portfolio Choice for Long-Term Investors", Oxford University Press.

Campbell, J.Y., Chan, Y.L., and Viceira, L.M. (2003), "A multivariate model of strategic asset allocation", Journal of Financial Economics 67, 41-80.

Cochrane, J.H. (2001), "Asset Pricing", Princeton University Press, New Jersey.

Cochrane, J.H. (2006), "Financial markets and the real economy", University of Chicago, December 2006 (forthcoming in "Financial Markets and the Real Economy", edited by Cochrane, J.H., Edward Elgar).

Cochrane, J.H. (2008), "The dog that did not bark: A defense of return predictability", Review of Financial Studies 21, 1533-1575.

Cochrane, J.H., and Hansen, L.P. (1992), "Asset pricing explorations for macroeconomics", NBER Macroeconomics Annual, 115-165.

Diebold, F.X., Ohanian, L.E., and Berkowitz, J. (1998), "Dynamic equilibrium economies: A framework for comparing models and data", Review of Economic Studies 65, 433-451.

Elliott, G., Rothenberg, T.J., and Stock, J.H. (1996), "Efficient tests for an autoregressive unit root", Econometrica 64, 813-836.

Engle, R.F., and Granger, C.W.J. (1987), "Co-integration and error-correction: Representation, estimation, and testing", Econometrica 55, 251-276.

Engsted, T. (2002), "Measures of fit for rational expectations models", Journal of Economic Surveys 16, 301-355.

Evans, G.W., and Honkapohja, S. (2005), "An interview with Thomas J. Sargent", Macroeconomic Dynamics 9, 561-583.

Fama, E.F., and French, K.R. (1988), "Dividend yields and expected stock returns", Journal of Financial Economics 22, 3-27.

Fama, E.F., and French, K.R. (1996), "Multi-factor explanations of asset pricing anomalies", Journal of Finance 47, 426-465.

Faust, J., and Whiteman, C.H. (1997), "General-to-specific procedures for fitting a data-admissible, theory-inspired, congruent, parsimonious, encompassing, weakly-exogenous, identified, structural model to the DGP: A translation and critique", Carnegie-Rochester Conference Series on Public Policy 47, 121-161.

Fernández-Villaverde, J., and Rubio-Ramírez, J.F. (2004), "Comparing dynamic equilibrium models to data: a Bayesian approach", Journal of Econometrics 123, 153-187.

Granger, C.W.J., and Newbold, P. (1974), "Spurious regression in econometrics", Journal of Econometrics 2, 111-120.

Haavelmo, T. (1944), "The probability approach in econometrics", Econometrica 12, Supplement: 1-115.

Hansen, L.P., and Jagannathan, R. (1991), "Implications of security market data for models of dynamic economies", Journal of Political Economy 99, 225-262.

Hansen, L.P., and Jagannathan, R. (1997), "Assessing specification errors in stochastic discount factor models", Journal of Finance 52, 557-590.

Hansen, L.P., and Sargent, T.J. (1980), "Formulating and estimating dynamic linear rational expectations models", Journal of Economic Dynamics and Control 2, 7-46.

Hansen, L.P., and Sargent, T.J. (2007), "Robustness", Princeton University Press.

Hansen, L.P., and Singleton, K.J. (1982), "Generalized instrumental variables estimation of nonlinear rational expectations models", Econometrica 50, 1269-1286.

Hendry, D.F. (1995), "Dynamic Econometrics", Oxford University Press.

Hoover, K.D., and Siegler, M.V. (2008a), "Sound and fury: McCloskey and significance testing in economics", Journal of Economic Methodology 15, 1-37.

Hoover, K.D., and Siegler, M.V. (2008b), "The rhetoric of 'Signifying nothing': a rejoinder to Ziliak and McCloskey", Journal of Economic Methodology 15, 57-68.

Kydland, F.E., and Prescott, E.C. (1982), "Time to build and aggregate fluctuations", Econometrica 50, 1345-1370.

Kydland, F.E., and Prescott, E.C. (1996), "The computational experiment: An econometric tool", Journal of Economic Perspectives 10, 69-85.

Leamer, E.E., (2004), "Are the roads read? Comments on "Size matters", Journal of Socio-Economics 33, 555-557.

Lewellen, J. (2004), "Predicting returns with financial ratios", Journal of Financial Economics 74, 209-235.

Lucas, R.E. (1987), "Models of Business Cycles", Blackwell, Oxford.

Lucas, R.E., and Sargent, T.J. (1981), "Rational Expectations and Econometric Practice", University of Minnesota Press, Minneapolis.

McCloskey, D.N. (1983), "The rhetoric of economics", Journal of Economic Literature 21, 481-517.

McCloskey, D.N, and Ziliak, S.T. (1996), "The standard error of regressions", Journal of Economic Literature 34, 97-114.

McCloskey, D.N, and Ziliak, S.T. (2008), "Signifying nothing: reply to Hoover and Siegler", Journal of Economic Methodology 15, 39-55.

Mehra, R., and Prescott, E.C. (1985), "The equity premium: A puzzle", Journal of Monetary Economics 15, 145-161.

Nelson, C.R., and Kim, M.J. (1993): "Predictable stock returns: The role of small sample bias", Journal of Finance 48, 641-661.

Phillips, P.C.B. (1986), "Understanding spurious regressions in econometrics", Journal of Econometrics 33, 311-340.

Schorfheide, F. (2000), "Loss-function based evaluation of DSGE models", Journal of Applied Econometrics 15, 645-670.

Spanos, A. (1986), "Statistical Foundations of Econometric Modelling", Cambridge University Press, Cambridge.

Stambaugh, R.F. (1999), "Predictive regressions", Journal of Financial Economics 54, 375-421.

Xu, Y. (2004), "Small levels of predictability and large economic gains", Journal of Empirical Finance 11, 247-275.

Watson, M.W. (1993), "Measures of fit for calibrated models", Journal of Political Economy 101, 1011-1041.

Ziliak, S.T., and McCloskey, D.N. (2004), "Size matters: The standard error of regressions in the American Economic Review", Journal of Socio-Economics 33, 527-546.

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

Research Papers 2009

CREATES
Center for Research in Econometric Analysis of Time Series

2009-02: Morten Ørregaard Nielsen: Nonparametric Cointegration Analysis of Fractional Systems With Unknown Integration Orders

2009-03: Andrés González, Kirstin Hubrich and Timo Teräsvirta: Forecasting inflation with gradual regime shifts and exogenous information

2009-4: Theis Lange: First and second order non-linear cointegration models

2009-5: Tim Bollerslev, Natalia Sizova and George Tauchen: Volatility in Equilibrium: Asymmetries and Dynamic Dependencies

2009-6: Anders Tolver Jensen and Theis Lange: On IGARCH and convergence of the QMLE for misspecified GARCH models

2009-7: Jeroen V.K. Rombouts and Lars Stentoft: Bayesian Option Pricing Using Mixed Normal Heteroskedasticity Models

2009-8: Torben B. Rasmussen: Jump Testing and the Speed of Market Adjustment

2009-9: Dennis Kristensen and Andrew Ang: Testing Conditional Factor Models

2009-10: José Fajardo and Ernesto Mordecki: Skewness Premium with Lévy Processes

2009-11: Lasse Bork: Estimating US Monetary Policy Shocks Using a Factor-Augmented Vector Autoregression: An EM Algorithm Approach

2009-12: Konstantinos Fokianos, Anders Rahbek and Dag Tjøstheim: Poisson Autoregression

2009-13: Peter Reinhard Hansen and Guillaume Horel: Quadratic Variation by Markov Chains

2009-14: Dennis Kristensen and Antonio Mele: Adding and Subtracting Black-Scholes: A New Approach to Approximating Derivative Prices in Continuous Time Models

2009-15: Charlotte Christiansen, Angelo Ranaldo and Paul Söderlind: The Time-Varying Systematic Risk of Carry Trade Strategies

2009-16: Ingmar Nolte and Valeri Voev: Least Squares Inference on Integrated Volatility and the Relationship between Efficient Prices and Noise

2009-17: Tom Engsted: Statistical vs. Economic Significance in Economics and Econometrics: Further comments on McCloskey & Ziliak