

Too much testing, poor testing: a review of research on
empirical economics

Alfonso Novales
Departamento de Economía Cuantitativa
Universidad Complutense
28023 Madrid

February 2006

Abstract

1. Introduction

For some time, a number of authors have criticized some aspects of econometric practice, specifically what seems to be a fallacious discussion of the economic relevance of explanatory variables in econometric models. This paper argues that there are additional fundamental aspects of the practice of statistical inference methods in economics which deserves strong criticism. Given the emphasis of applied econometrics research on classical hypothesis testing, most of the issues discussed in the paper deal with the way how testing is implemented and the implied results are interpreted. In particular, applied research in economics centers around running a set of tests for whether potential explanatory variables contain relevant information on the phenomenon which is being analyzed, so we will be discussing several aspects of so-called *significance tests*.

These have been object of criticism by previous others, like McCloskey and Ziliach [see Appendix], who made a detailed analysis of some aspects of statistical inference in empirical papers published during the eighties in American Economic Review. They emphasized important issues mostly related to significance tests, but their criticism could be easily extended to other areas. In the last decades there has been a clear progress in departing from standard linear models, which can be convincingly argued to be too limited to accommodate the complexity of relationships among economic variables. Unfortunately, even though more interesting models are increasingly being used, application of statistical inference methods is still too mechanical, and that is a frequent source of flaws when interpreting estimated models. In fact, McCloskey and Ziliach repeated their evaluation of empirical AER papers over the nineties, finding no evidence of improvement on the checkpoints of their tests. We will describe the main elements of the practice of statistical inference that should be amended, and will provide some examples to illustrate the main issues. Even though we will make references to "the dependent variable" as well as to set of "explanatory variables" or to least-squares estimation, our criticism on the application of statistical inference methods in Economics may apply to all kinds of models and estimation methods.

2. The limitations: Economics is a nonexperimental science

2.1. Too much testing

As other social sciences, Economics is a non-experimental science. This is explicitly recognized in most textbooks on statistical methods written for Economics and Business and yet, it is generally forgotten when doing research. Because of being nonexperimental, a single sample is available in most cases on the variables of interest, which limits the strength of the conclusions that can be reached by standard statistical methods. A striking example is the emphasis on unbiasedness in most Econometrics textbooks, which is not justified by itself unless we could replicate the sample by repeated experimentation in the same conditions and average the numerical values of the estimator across samples. Since we will only see a single realization (a sample of size one) from the distribution of the estimator, unbiasedness is only interesting if it comes together with a relatively low variance. This is an example of the statements made in the introduction: even though minimizing the mean squared error has sometimes been suggested as a reasonable

criterion for choosing among alternative estimators, most economists would still disregard biased estimators if they had to select one among a set of alternative estimation methods.

Another implication of the non-experimental character of Economics is that, contrary to experimental sciences, the researcher does not control the values taken by the explanatory variables in a model. As a consequence, with the exception of methodologies which deal with all variables as being endogenous, as VAR models, the stochastic character of explanatory variables should always be recognized. But the standard use of t - and F -statistics in small samples rests in rather restrictive assumptions, from the nonstochastic character of explanatory variables, to the Normality of the error term, and parameter stability (both, for coefficients and for the covariance matrix).

Therefore, a more careful use of standard distribution theory when testing hypothesis in small samples is warranted. A more extended use of appropriate estimation methods, like instrumental variables or generalized method of moments, may be more appropriate, although it will still be subject to the consideration of possible parameter variation. Granting parameter stability, we in fact must rely on asymptotic theory, whose small sample properties are usually unknown. Large samples would solve this issue, although we have the discussion between more data points or a longer sample period, but parameter stability may then be less acceptable in the second case. Consistency should be preferred to unbiasedness as an estimation characteristic, and parameter variations should be a central part of any estimation research, even when using cross-section data.

A further nontrivial implication of the fact that the inputs in a relationship model are not controlled by the researcher, is that they will generally contain overlapping information, which precludes the standard direct interpretation of estimated coefficients, as we will discuss below. The common information in explanatory variables that we know as collinearity also affects the results of standard hypothesis testing. The variance of the least-squares estimator of a single coefficient increases without bound with the degree by which the associated variable is explained by the remaining explanatory variables. Given a numerical estimate for that coefficient, the t -statistic will become arbitrarily small as collinearity increases, yielding the appearance that the associated variable does not explain the dependent variable.

Precisely because of the strong limitations described above, it is surprising how much emphasis is placed in formally testing competing theories through parameter restrictions they imply on estimated models.

2.2. Poor testing

A first implication of the preliminary observations above is that a detailed data analysis could be a good substitute for systematic hypothesis testing. When implemented, testing results should at least be interpreted with caution. Additional reasons for caution are:

1. The loss of power that arises by the fact that, unavoidably, it is a short set of parameter implications, and not the whole theory, which is put to test, is far from irrelevant. There may be an alternative theory implying the same hypothesis being tested. Furthermore, not rejecting the null hypothesis does not say anything about the evidence concerning additional implications from theory.
2. Parametric restrictions are tested at widely agreed upon significance levels which are not

justified on the basis of the amount of data information available. Two issues show up here: *i*) the sample information is useful only if it produces precise estimates, so confidence levels should be positively related to sample size or precision. It is relatively easy to reject any null hypothesis at all with a large enough sample, *ii*) the probability of a type I error is determined by the significance level chosen, which in turn conditions the power of the test and hence the probability of a type II error. A loss function should be used to weight the two possibilities, and a significance level (or test size) should be chosen to minimize expected loss. When possible, the sample size can be adjusted to improve the testing situation.

3. Economists seem to have forgotten about the fact that rejecting the null in a classical hypothesis testing approach requires of two conditions: *i*) that the available sample evidence be contrary to the null hypothesis, *ii*) while being favorable to the alternative hypothesis. This is specially relevant in one-side hypothesis testing, which is important because, contrary to the impression one gets by reading empirical research on economics, economists often have enough information to make two-sided tests inappropriate.
4. An even more important problem is the extended interpretation that lack of rejection of the null hypothesis amounts to *having found conclusive evidence* that the hypothesis is true. This is a consequence of a more serious flaw, the total absence of statistical power functions from applied economic research.
5. Given the high frequency of cases in which formal statistical testing in economics is of the form of significance tests, the second big fault in applied statistical inference in economics is the *identification of two very different concepts: statistical significance and information content*. The point is that *t*-statistics do not have much to say on the information content of a explanatory variable.
6. Even when this is fully understood, a final remark relates to the fact that, by nature, the analysis on the explanatory power or information content of a given variable is always *conditional* on the other variables included in the model.
7. Information content in a variable should be analyzed in models that use it as its only explanatory factor. Sure enough, the estimated coefficient in such a model is a biased estimate of the impact of a unit change in that variable leaving other explanatory variables unchanged. However, it may be a possibly unbiased estimate of the global impact of a unit change in that variable taking into account associated changes in additional explanatory variables. And it is the latter, rather than the former, what should be of interest to the economist.

These initial thoughts should already suggest that the situation faced by applied economists should lead to a cautious interpretation of results, far from the statements we often read like "*having shown that ...*", "*having shown conclusive evidence....*".

There are ways out. Researchers should follow a determined strategy to evaluate information content having statistical significance test as one, but not the only one, approach followed to reach a conclusion. Sound research requires that the question under analysis be examined from a variety of different empirical methodological perspectives. They will produce results which

may well be less than fully coincident, and it is the role of the researcher to evaluate which of the methodologies is more reliable in each application, and base on them the choice of specific conclusions. In any event, the researcher should always report all the information generated in his analysis, so that any reader may reach his own conclusions. On the contrary, the standard practice of reducing all the sample information regarding a given hypothesis to a few figures at most, the numerical value of the test statistics used, is a sound strategy. We will return to this issue of information reduction in a later section.

3. Measuring the information content on an explanatory variable

Leaving aside models designed only for forecasting purposes, formal hypothesis testing usually follows the standard stages of model specification and estimation. Most of the comments we make on testing can be described in reference to the many theoretical situations in which a unit elasticity is tested for, or in relation to significance tests for individual coefficients. We devote this section to reviewing the main points for contention regarding hypothesis testing.

3.1. Ignoring the power function of a test

Parametric restrictions are tested at widely agreed upon significance levels which are not justified on the basis of the amount of data information available. For statistical purposes, it is not the length of the sample, but the amount of information, which is relevant. And the quantity of information contained in a sample is directly related to the precision of the estimates it produces. At the outset of any econometrics course, when discussing least-squares estimation of a simple linear regression model, it is explained how a long sample in which an explanatory variable barely changes amounts to little information, which will be reflected in the fact that the estimator variance will be large. This is one of those simple, but illuminating results, that should be emphasized. The search for good sample information is justified by the need to reach enough precision in estimation which, in turn, will lead to powerful tests, which should be the goal of the applied researcher interested in formal hypothesis testing. It is important to bear in mind the connection between the precision situation reached in estimation, and the power of classical hypothesis tests based on those estimates.

But the usual testing practice in economics is subject to serious difficulties. One of the most important problems is the widespread interpretation that lack of rejection of the null hypothesis amounts to *having found evidence* that the null hypothesis is true. This is a profound pitfall that emerges from the fact that applied economists have forgotten about the existence of the power function of a test¹. We seem happy with a widespread use of the p -value of the test but, unfortunately, that is not enough. The p -value informs us of the probability of finding evidence more contrary to the null hypothesis under consideration than that contained in the available sample, provided H_0 is true. A low p -value, below .05 say, is interpreted as suggesting that under H_0 , almost any other possible similar sample would have thrown more favourable evidence in favor of H_0 , so H_0 should be rejected as not being true. A large p -value would suggest that any similar sample, extracted under H_0 , would have been less favourable to the null. So, our sample is actually quite favourable, and H_0 should not be rejected.

¹See points 7 and 8 in D. Mc Closkey and S.T. Ziliak (Appendix)

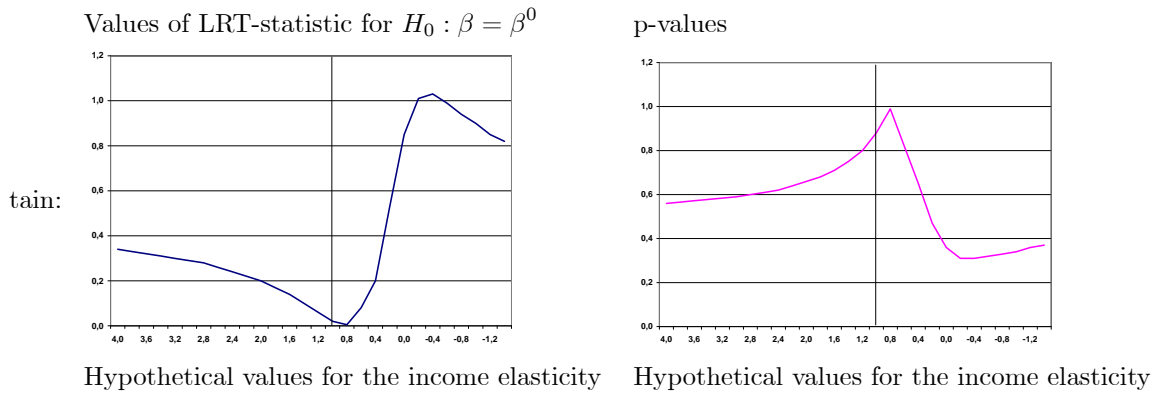
The p -value does not tell us much about the degree to which the sample evidence is contrary to other alternative null hypotheses. It could easily be the case that a p -value for a given null hypothesis is of .80, say, while not being much lower for a wide range of values around those defining the null hypothesis. So, the p -value is hardly a substitute for a power function.

Not examining the likelihood of alternative null hypothesis is the basis for acting as if not rejecting a null hypothesis would amount to having it shown to be true. Not examining the power function of the hypothesis being tested is possibly one of the most important failures of hypothesis testing in applied economics. The power function quantifies in terms of probability the amount of evidence against null hypothesis different from the one being put to a formal test. This analysis should be central in any exercise on hypothesis testing. No test should be run without a more or less formal analysis of its power for local alternatives. In fact, it is hard to disregard the Bayesian view of approaching the confrontation of a given hypothesis with the data as an exercise in evaluating the likelihood of that hypothesis as well as local variations around it. Likelihood graphs and power functions should be displayed, at least when working in relatively simple frameworks.

The researcher should never place much emphasis on a null hypothesis that has not been rejected, if the power function reveals that alternative hypothesis in a relatively large neighborhood would not be rejected either, since that would reflect a low power of the test. Reduced power will usually come from low precision estimates, which is a main reason why searching for a relatively efficient estimator is important. Overparameterization is a leading cause for loss of efficiency and hence, loss of precision in estimation and loss of power in testing hypothesis on estimated economic models. A second cause is parameter variation, to which we will return later on.

3.1.1. An example: testing for a unit elasticity in the demand for money

As an example, Muscatelli and Hurn (1992) and Ericsson, Hendry and Trand (1994), who used the then relatively new cointegration approach to model the demand for money in the UK and tests the null hypothesis of a unit income elasticity. This is a very relevant characteristic when using the demand for money to implement monetary policy, since it conditions by how much the money supply should increase in coherence with given targets or forecasts for economic growth and inflation. [**Parameter instability**]. After showing how cointegration methods allow for capturing a new process of parameter instability in the demand for money due to a new round of financial innovation, the unit elasticity of income was not rejected and then, taken as true for monetary policy implementation. Garcia-Ferrer and Novales (19xx) follow up on this work, to ob-



The graph to the left displays the values of the likelihood ratio test statistic (LRT) for the hypothesis $H_0 : \beta = \beta^0$, β^0 being each of the numbers on the $(-2.5, 4.0)$ range, as shown in the horizontal axis. Notice that negative values are to the right and positive values to the left. The statistic gives us an indication of how contrary is the sample evidence with respect to each parameter value in the indicated range, in relative terms. That way, we see that there is much more evidence against negative than against positive elasticity income values. This is to be desired, from a theoretical point of view. The available data seems to discriminate well against zero or negative elasticities, as reflected in the sudden raise of values of the statistic. On the other hand, the evidence against positive values is less strong, and it does not discriminate well over values above 2.0. It takes its minimum value around 0.8, the maximum likelihood estimate. The researcher could now select a level of the LRT statistic, like 0.1 for instance, and characterize over the graph the interval of parameter values, with a value of the statistic below that level, $(0.6, 1.2)$, say. This seems as a good alternative to thinking in terms of confidence intervals.

The graph to the right shows p -values for the mentioned null hypothesis. Some information is common with the previous graph, since the p -value should be an inverse function of the value of the LRT -statistic. So, it is not surprising that the maximum p -value is of 1.0 for the maximum likelihood estimate, and that p -values are lower for negative than for positive elasticities. What is striking in this example is the fact that p -values stay above 0.30 for the range of parameter values considered, which is much wider than could be accepted from the point of view of monetary theory. In summary, the conclusion reached by the authors of a unit income elasticity should be downplayed. A quite more exact conclusion would be that, far from "having shown" a unit elasticity, at standard significance levels, the data cannot discriminate between a wide range of income elasticity levels.

Note: It is striking that so little discussion has been made of the standard practice of reducing all the information available in the sample regarding a given null hypothesis to a single number.

3.2. Misusing the t -statistic

Economists have also forgotten about the fact that rejecting the null in a classical hypothesis testing approach requires of two conditions: *i*) that the available sample evidence be contrary

to the null, *ii*) while being favorable to the alternative hypothesis. This is relevant in one-side hypothesis tests. But it is usually the case that we know that if a unit elasticity is not equal to one, then it must be less than one, or that if a given coefficient turns out to be statistically significant, then it must take a positive value. All these should be treated as one-sided tests, which makes a difference in terms of critical values. It might even be preferable not to test at all, but if we do, we should design tests right.

Let us consider the frequent case of testing for a unit coefficient in an estimated relationship, against the alternative that the coefficient be less than one. Let us assume that the estimated coefficient turns out to be a surprisingly large value like 1.85, with a standard deviation of .30. It may surprise some readers, but we are used to conclude that the null hypothesis should not be rejected. The conclusion may be flawless from the point of view of classical statistical theory, but statistics does not have much to do with the way how test results should be interpreted. In fact, statistics is far from being a set of rules to be mechanically applied. The decision should make the researcher worry about the reasons why such high estimate may have turned out. It may be the case that the sample is not fully appropriate, or it might even be the case that the researcher may want to change his view on the basis of the numerical estimate.

Let us now suppose that the standard deviation had been of .70. According to classical hypothesis testing theory, the null hypothesis would then not be rejected in a one- or in a two-sided test. The question here would be different, since a standard 95% confidence interval would include approximately the range of values (.45,2.55), which should appear as being too wide for most purposes. Here, the sample information has not allowed us to reach enough *precision* to make any sensible statement on the true value of the estimated parameter. It would be a very dubious conclusion to say that the sample information supports the unit value for that coefficient and yet, that is often the way how a result of this kind is reported.

3.3. Precision in estimation and test power

Here is where we see the inverse relationship between precision of an estimator and the power of testing a null hypothesis on that parameter. Low precision amounts to a large standard deviation and hence, large confidence regions. Most confidence regions will tend to include a large number of sensible null hypothesis, which would hence not be rejected, implying a low power for almost any test.

The mechanical use of the *t*-statistic is subject to type-I and type-II errors. Type-II errors arise because of reduced power. The *t*-statistic will take a low value when numerical estimates satisfy the null hypothesis to a good numerical approximation. But, it will also be small when parameters involved in the null hypothesis are estimated with reduced precision, even if estimates do not satisfy the null hypothesis with enough approximation. In the first case, not rejecting the null hypothesis will be a correct decision, being wrong in the second case. The latter is the low-power situation, in which null hypothesis different from the one considered would not be rejected either. The researcher should always analyze that. In most cases, the researcher must ponder by how much the null hypothesis is not satisfied by the numerical estimates. For that evaluation, the standard deviation is used as the unit of measure, and it could be too large a unit, leading to not rejecting the null hypothesis. There can be no absolute rules for that decision. The answer to the question of by how much should the estimates differ from the null hypothesis in order to reject it can only be given by the researcher on a specific application basis. The

analysis we suggested above on the money demand function, about choosing a percent change in the objective function, be this the sum of squared residuals or the likelihood function, and computing the region of the parameter space producing a value of the function inside that region, seems a sensible alternative.

3.3.1. An example: testing for the Expectations Hypothesis

A typical example arises in tests of the expectations hypothesis of the term-structure, or similar, like purchasing power parity tests, tests of the forecasting ability of futures prices, and so on. The researcher has available a cross section on which the hypothesis can be tested: interest rates at different maturities, different currencies, etc.. Estimates for the relationship model that tries to explain short term rates by lagged forward rates are presented as,

Projection: Future spot rate on lagged forward rate

$$H_0 : \beta = 1$$

Plazo	Beta	Desviación típica	R2	Estadístico t H0: Beta = 1
1 mes	0,96	0,04	0,54	-1,00
3 meses	0,93	0,06	0,42	-1,17
6 meses	1,10	0,15	0,32	0,67
1 año	1,40	0,24	0,22	1,67
3 años	1,63	0,47	0,15	1,34
5 años	2,40	0,85	0,04	1,65

and the hypothesis is evaluated. A mechanical, standard application of t -tests would lead us to conclude on the legitimacy of the Expectations hypothesis.

This is an example of testing under low precision. Estimated betas increase with maturity, and depart significantly from the unit value in the null hypothesis for maturities above one year. However, precision is also increasingly lower, as reflected in estimated standard deviations, with the result that t -statistics remain below the standard critical level for usual significance levels and most sample sizes. It is however, questionable that a strong emphasis on the Expectations hypothesis should be placed. With the scant information in the table, we could say that the hypothesis seems to hold reasonably well only for maturities under one year. The point is that with the reported estimated values, the hypothesis should not be tested for longer maturities².

It is important to realize that the null hypothesis of a unit coefficient would not be rejected even if the test was one-sided and the precision was much higher than reflected in the table above. The reason for that would be the fact that, while being contrary to the theoretical value defining in the null hypothesis, the numerical estimates are not favorable to the alternative hypothesis of $H_0 : \beta < 1$. That was the point of the example in the previous section.

The literature on the Expectations Hypothesis and similar theoretical propositions is a good example for other unfortunate practices. One is the striking identification between good adjustment and forecasting ability. Even since the first papers by Fama, forward rates are said to be good predictors of future short term spot rates whenever the Expectations Hypothesis is not rejected, or as soon as some explanatory power is detected without need of any hypothesis testing, even though *no formal prediction exercise is ever performed* [Dominguez and Novales (19xx)].

²Which would an example of what Mosteller and Tukey (19xx) call *concealed inference*.

4. Significance tests

Our previous comments on low precision in estimation leading to low power in testing apply in full length to significance tests, the statistical exercise practiced most often in empirical research in Economics. Usual causes for reduced precision will be scant sample information, *i.e.*, little variation in explanatory variables, excessive overlap of information among explanatory variables, parameter variation, nonlinearities in cross-section data not incorporated in the model, a low signal-to-noise ratio. The point is that all them can be compatible with a more than acceptable information content in the variable whose coefficient is undergoing a significance test.

4.1. Statistical significance and information content

Accepted as it is in practice, the standard identification between lack of statistical significance of a given coefficient and lack of information content in the associated variable is profoundly falacious. On the one hand, one is a statistical property of a given *coefficient*, while the other is a characteristic of the relationship between two variables. As pointed out in previous work³ applied research in economics focuses on the first concept, when it is the second one that should be of interest. The point is that it is relatively easy to have one of them without the other. As already mentioned, in any situation in which *precision* of estimates is not too large, parameter regions for standard confidence levels will be *large*, and the numerical value of associated test statistics will be small. In the case of testing for the significance of a single coefficient, lack of precision implies a large standard error of the estimate coefficient and hence, to a small value for the *t*-statistic, inducing the test not to reject the null hypothesis of lack of significance of *the associated variable*.

Economists should be sensible towards the possible causes of reduced precision, and be aware in their specific research that such is not the main reason why the null hypothesis is not rejected. Those who perceive economic relationships as subject to smooth, gradual changes in intensity, need to be aware that time variation is one of the main reasons to lead to low precision estimates of parameters incorrectly assumed to be time-invariant. In this type of situations, which may be safely deemed to be not too infrequent in economics, mechanical application of classical hypothesis testing will be biased towards concluding the lack of relevance of the associated variable. A purely technical question, in this case related to a time-stability assumption on the parameter, together with the confusion between the two concepts, lack of statistical significance and lack of economic relevance, leads to this wrong conclusion.

In the specific case of significance tests, a type-I error would lead us to conclude that the associated variable has some information content regarding the dependent variable. In the case of a small coefficient estimated with high precision, the conclusion is dubious. The statistical significance of the coefficient would be unquestionable, but the identification between this statistical property of an estimated coefficient and the concept of information content of the associated variable would lead to a wrong conclusion⁴. A type-II error arises from low precision estimates. We would then conclude on the lack of explanatory power for the explanatory variable when it

³McCloskey and Ziliak (1996) and (2004)

⁴As we will discuss below, it is simple to anticipate that the expected variation in the explanatory variable should play a role, together with the estimated numerical value of the coefficient and the acceptable range of values for it (confidence region), which should determine the potential relevance of the associated variable. And once again, colinearity will also be a factor.

could actually contain relevant information. We must remember that in the presence of time variation, any estimation procedure will give us some sort of average of the true values taken by the coefficient. Precision will depend on the type of time variation, being lowest when the true value of the coefficient has increased or decreased all along the sample.

Among all practical pitfalls, I believe that wrongly concluding on the lack of information content of a given explanatory variable because of the described scenario is the most widespread one. Needless to say, the practice of sign-econometrics should be eliminated. Statements like: "... *the coefficient has the wrong sign, although it is not statistically significant.*", [what MacCloskey and Ziliach call "sign econometrics") or "... *it has the sign implied by the theoretical model, but it does not explain Y ...*" lack rigor⁵.

Of course, the whole point is how could we conclude on the information content of the explanatory variable, which is discussed below.

4.2. The test for information content is a conditional test

A second line of difficulties arises because we forget about the important fact that significance tests only test for whether *the variable under consideration adds some relevant information to that provided by variables already contained in the model*. When, by following standard practice, lack of statistical significance in a coefficient takes the researcher to conclude that the associated variable is not relevant to explain changes in the dependent variable, he is adding to the confusion of the two concepts, a mistaken conclusion. What he could say is, if anything, that the referred variable does not contain information on Y additional to that already contained in the vector of explanatory variables Z in the model.

In fact, standard F -tests of significance compare residuals from two models, both including the rest of the variables Z in the model, one of them additionally including and the other model excluding the variable X undergoing the test. Hence, it should be clear that the additional information hypothesis is the one being tested. We have to bear in mind that to conclude on this, collinearity is central. Consider two highly correlated variables X and Z , both having a similar, high correlation with the dependent variable Y . Neither one of them will have a statistically significant coefficient when added to a model that explains Y using the other explanatory variable. However, each one of them will generally have a statistically significant coefficient when used by itself to explain Y . There is nothing paradoxical about this result, since the two models: one using X as the only explanatory variable, and the other including both, X and Z , are asking different questions about the information content in X .

In fact, it is unfortunate that standard practice finds convenient to summarize the differences between both sets of residuals to a single figure, rather than following alternatives lines of reasoning to compare sets of residuals from different models as described below. We will get back to this issue later on, but let us now advance that the information content in an explanatory variable cannot be compared between alternative specification models containing different sets of explanatory variables. The information content of X on Y is not an absolute concept, depending rather on the model specification and the sample used.

⁵See points 10 and 11 in Mc Closkey and Ziliak.

5. Interpreting estimated models

The nonexperimental character of economics precludes the design of samples obtained in experiments where perfectly controllable inputs are designed to be independent by construction. As a consequence, given the extensive simultaneity in most economic areas, makes colinearity to be a standard situation in an estimated model relating economic variables. As it is the case with other issues like autocorrelation and heteroskedasticity, there is some sense in which colinearity should not be thought as a potential difficulty, rather than as an always present situation that conditions the way estimated models should be interpreted.

Estimated models are still often interpreted on an individual coefficient basis. The researcher takes each individual coefficient and examines, *i*) statistical significance through the numerical value of its t -statistic, *ii*) interprets how much influence the associated variable has on the dependent variable by the numerical value of the estimated coefficient. The first practice tends to be misleading because of the identification between *statistical significance* and *informational content* already mentioned.

The second practice is striking, for several reasons. A general idea is that it is questionable that the value of the coefficient provides useful information on the relevance or strength of the association between explanatory and dependent variables, as we will discuss below. An additional reason for making this a dubious practice is the colinearity among explanatory variables. Econometrics textbooks usually devote a chapter to multicollinearity, where the problem of the possible singularity of the $X'X$ matrix is discussed, and potential solutions, in the form of excluding variables, restricting the set of explanatory variables, etc., are proposed. This is of little value in practice. With actual data, exact collinearity arises infrequently, because that would need of any two explanatory variables being of very similar nature or taking values very highly correlated over the sample. That will usually be due to a specification problem, easy to correct.

The problem with collinearity is not the singularity of the $X'X$ matrix, but that it forces the researcher to be careful when interpreting an estimated model.

5.1. Measuring the relevance of a explanatory variable: analyzing the residuals

Statistical testing reduces the sample information regarding the null hypothesis to a single figure. Alternative strategies are needed to move away from the difficulties mentioned in the previous sections. Maybe the most important one is to recover the role of the residuals in model evaluation. No matter what estimation method is used, the dependent variable can always be decomposed between adjusted values and residuals. Residuals have the same size as the sample, and contain detailed information that should not be ignored. The following consideration should be convincing enough: let us assume that a given variable has been quite influential on the dependent variable, but also that this has been the case for a subsample amounting to a relatively small percentage of sample data. The single figure summarizing the information for the whole sample may well miss this effect, leading the researcher to conclude on the irrelevance of the explanatory variable.

Alternatively, let us consider a comparison between two sets of residuals, those obtained estimating models that include and exclude the variable X under consideration. Both models should include as explanatory the vector Z made up by the other explanatory variables. In its simpler version, this comparison could take the form of a scatter diagram between both sets of

residuals. If the points in the diagram turn out to be aligned along the 45° degrees line, that would indicate that the two sets of residuals are the same. The right conclusion then is that X does not contain information on Y *additional* to that already contained in vector Z .

A short-lived effect of X on Y would show in that diagram as a set of points apart from the mainstream, which would be aligned along the diagonal. The researcher could then look for patterns characterizing that set of points: do they correspond to a specific point in time? are they associated to negative, positive, large or small values of X or Y ? That way, omitted effects could also be detected.

Sure enough, more formal procedures for comparing the two sets of residuals are available. Particularly interesting seem non-parametric tests, which do not require Normality of the residuals, parameter stability, lack of heteroskedasticity, and so on. However, tests comparing the frequency distributions of residuals like Kolmogorov-Smirnov or Pearson-like tests would not be appropriate. Similar distributions could mask very noticeable differences between the two sets. For instance, imagine that the model that does not include X has low residuals at the beginning of the sample, and larger residuals, the model including X as explanatory variable doing the opposite. The two sets of residuals could be wildly different and still produce similar histograms. The type of tests needed are those comparing residuals for the same observation: sign test, signed rank test, Wilcoxon test, and so on.

5.2. Interpreting individual coefficients

The coefficient associated to an explanatory variable X measures the size of the effect on the dependent variable of a unit change in X , all other explanatory variable remaining unchanged. However, this characteristic is likely to be of little consequence in most cases for two reasons: *i*) that colinearity makes unlikely that all other explanatory variable can remain unchanged when one of them changes, *ii*) that a unit change may not be representative of the size of a likely change in X . Consequently, *comparing the relative importance of two explanatory variables on the basis of their estimated coefficients is inappropriate*. From our previous discussion on how the value of the t-statistic gets conditioned by the precision of estimates, the reader should infer that *comparing the relative importance of two explanatory variables on the basis of the t-statistics for their associated coefficients is also rather inappropriate*. Sometimes, *the relevance of an explanatory variable in two subsamples* is also judged on this basis, being subject to the same criticism.

5.2.1. Measuring the quantitative impact

In time series as well as cross-section data, the standard deviation of a variable indicates the degree of variability over the sample. In stationary time series data, the standard deviation indicates the average period-to-period change in that variable. If we want to quantify the effect of such an X on the dependent variable, then the product $\beta\sigma_X$ would give us an indication of the period-by-period influence of X on Y , and it would seem as an statistic that could help to compare the relative importance of two explanatory variables. It shows that the explanatory power of X on Y depends not only on its coefficient, but also on the sample volatility of X . This analysis is in line with Tinbergen's suggestion to display in time series samples the lines for βX_t . The latter is potentially more interesting because it would be able to capture periods

of different quantitative relevance for X . Unfortunately, it is subject to the limitations emerging from colinearity among explanatory variables.

5.2.2. An example: Geographical dispersion of firm's exports

In an interesting empirical paper in *Revista de Economía Aplicada* (2003), a sample of Spanish firms is used to characterize the main determinants of geographic dispersion of exports. Potential explanatory variables used are the number of employees, the number of products, whether the firm is participated by foreign capital (dummy variable), the import ratio, an index of industrial concentration, ratios for expenditures in research and development and publicity, and an indicator of provincial concentration in production. The index of geographical dispersion of exports to be explained takes values in (0,1). Maximum likelihood estimates presented in the paper are shown in the table in the "Beta"-row. Some descriptive statistics included in de paper are also shown in the Table under the row labels: "mean", "Maximum", "minimum" and "standard deviation". The standard use of the t -statistics leads the authors to conclude on the relevance of all the variables in the table to explain the index of geographic dispersion of exports, the evidence being weaker in relation to the import ratio variable and the index of industrial concentration.

	Indice de dispersión geográfica de exportaciones	Empleo <50	Empleo (50,100)	Empleo (100,200)	Empleo >200	Segundas líneas productos	Participación capital extranjero	Ratio importador	Concent. industrial	I+D/Ventas	Publicidad/Ventas	Concent. provincial
Medias muestrales												
Media	0,243	0,203	0,349	0,221	0,226	4,375	0,094	0,114	0,185	0,005	0,009	0,175
Beta			0,026	0,057	0,099	0,004	0,047	-0,015	-0,006	0,051	1,302	0,087
			(4,29)	(7,25)	(13,2)	(2,56)	(6,53)	(1,96)	(1,54)	(2,35)	(6,44)	(6,79)
Producto						0,018		-0,002	-0,001	0,000	0,012	0,015
Mínimo	0	0	0	0	0	0	0	0	0,036	0,001	0,001	0,052
Máximo	0,560	1	1	1	1	82	1	0,724	0,810	0,083	0,086	0,745
Producto			0,026	0,057	0,099	0,328	0,047	-0,011	-0,005	0,004	0,111	0,060
Desviaciones típicas												
	0,170					5,480		0,200	0,140	0,010	0,060	0,180
						0,022		-0,003	-0,001	0,001	0,078	0,016

Even without examining in detail de sample, descriptive statistics provided in the paper allows for a more interesting analysis of information content than the one emerging from exclusive use of t -statistics. The product of sample averages time estimated coefficients gives us an indication of the level of the average effect of each explanatory. Still more interesting are the cross products of the estimated coefficient times the sample range of observed values for each explanatory variable, or their standard deviations, as explained above. The latter suggests the importance of publicity expenditures, with a $\beta\sigma_X$ value close to half of the standard deviation in the dependent variable. Being an indicator of average change between sample observations, this standard deviation is what we would like to explain. Differences in the number of product lines explain 17% of the standard deviation of the dispersion index, with provincial concentration explaining an additional 9%.

Standard deviations are clearly inappropriate for dummy variables, indicated in yellow in the table. For them, as for any other explanatory variable, we can examine the product of coefficients times sample range. This analysis suggests the importance of the number of product lines, for which this impact estimate is 58% of the range observed for the dependent variable, while confirming the relevance of publicity, which can explain 20% of the sample range for the dispersion index. The index tends to increase with size, the difference between a firm with more

than 200 workers and a small one explaining 28% of the range for the dispersion index. The difference between participated and non-participated firms is of .047, or about 7% of the average variation in the dispersion index. Notice that these numbers should not be expected to add up to 100%, because of the collinearity among the explanatory variables.

This analysis would not concede significant explanatory power to the import ratio, industrial concentration or RD-expenditures.

5.2.3. Colinearity

Dealing with colinearity is not easy, and it cannot be avoided, since more often than not, explanatory variables display a nontrivial degree of association. We can borrow from the VAR literature that computes impulse responses, i.e., dynamic effects of each variable on all others by first orthogonalizing the model innovations. In a model with dependent and explanatory variables, we can establish in each particular application a ranking of relevance among explanatory variables. This ranking can be made on the basis of theoretical reasoning, or by examining one-to-one relationships with the dependent variable.

To compare the relevance of explanatory variables X and Z to explain Y , we could do as follows: *i*) obtain residuals \hat{u}_z from a model explaining Z by X . If we think of a least-squares projection, that residual will have zero correlation with X , by construction. The residual \hat{u}_z could then be safely interpreted as the component of Z that cannot be explained by X , since it would not have information in common with that variable, *ii*) add that residual to the model that explains Y using X as its only explanatory variable. We would then in fact be asking whether the information in Z which is not common to X contains any explanatory power on Y additional to that already contained in X . That seems as a sensible question to ask. The exercise can be repeated by obtaining the residual \hat{u}_X from a projection of X on Z , to be added to a model that explains Y by Z alone.

It is clear that the three pairs (X, Z) , (X, \hat{u}_z) , (\hat{u}_X, Z) contain the same information on Y , and the fit of the models having each of these pairs should be the same. The difference is that the explanatory power is split into the two variables with no overlapping in the case of the (X, \hat{u}_z) , (\hat{u}_X, Z) -pairs, and with some overlap in the case of the (X, Z) -pair. The variable whose residual contributes the least when added to the model explaining Y , should be considered the less relevant variable in the (X, Z) -pair.

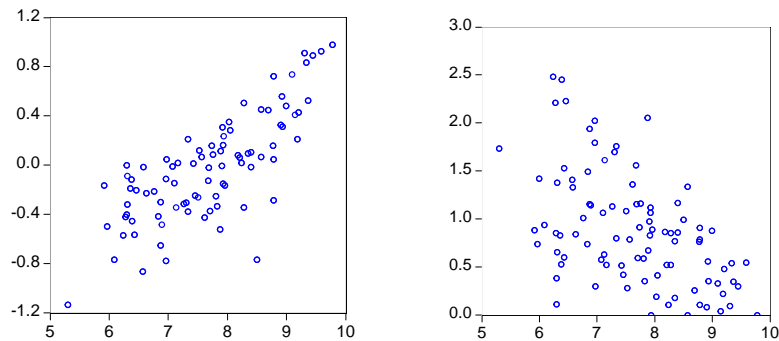
This strategy can be applied even if we are questioning the relative importance of X and Z conditional on a vector W of explanatory variables already included in the model.

5.3. Comparing the explanatory power of two variables

5.3.1. An example: Macroeconomic versus institutional factors for growth

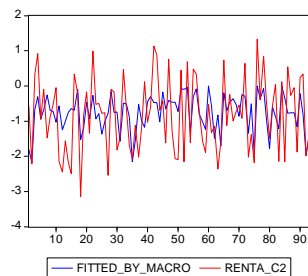
As an example, consider work by C. Sebastián [Reference], where the author attempts to evaluate whether it is macroeconomic factors or institutional factors the ones which are more responsible for growth. Without entering into details which would be out of place in this report, let us just mention that the author uses an average of macroeconomic policy indicators ("macro"), as well as a synthetic indicator of institutional development ("institutions") to explain growth over a given sample period in a sample of 93 countries. Both indicators show a linear correlation coefficient of -.55, so colinearity is not too large, but high enough to make us worry about

interpretation of individual coefficients. Scatter diagrams between each explanatory variable and the dependent variable are,

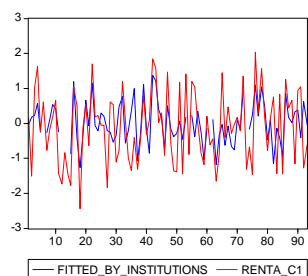


Correlation coefficient between "institutions" and "income" is of .76, being of -.51 between the "macro" indicator and the income variable. The first regression shows the fit provided by the use of the macroeconomic indicator as well as the component of "institutions" not explained by "macro", while the second regression uses as explanatory variables "macro", together with the component of "institutions" not explained by "macro". As expected, the goodness of fit of both models is the same, and identical to the one obtained with both indicators. Graphs to the right display the fit obtained from using in each regression the variable which is used in original form. A simple inspection suggests that institutional factors explain growth better. Correlation coefficient between the original income variable and the version fitted by "macro" is of .51, while correlation with the version fitted by "institutions" is of .58.

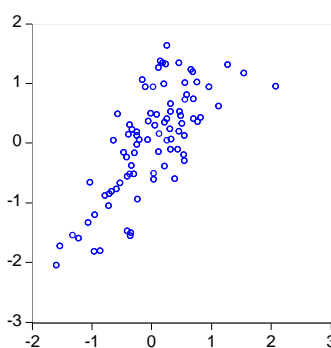
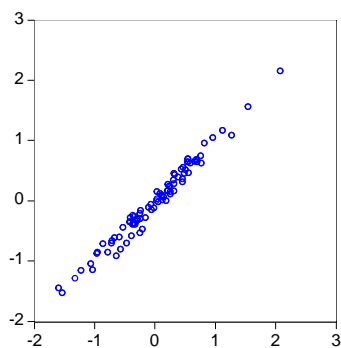
Dependent Variable: RENTA		Sample: 1 93		
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	8.445	0.125	67.285	0.00
MACRO	-0.879	0.119	-7.328	0.00
INSTITUTIONS_MACRO	1.662	0.200	8.306	0.00
R-squared	0.593	Mean dependent var	7.687	
Adjusted R-squared	0.583	S.D. dependent var	1.029	
S.E. of regression	0.663	Akaike info criterion	2.052	



Dependent Variable: RENTASample: 1 93				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	7.734	0.071	108.48	0.00
INSTITUTIONS	1.825	0.166	10.98	0.00
MACRO_INSTITUTIONS	-0.211	0.144	-1.46	0.14
R-squared	0.593	Mean dependent var	7.687	
Adjusted R-squared	0.583	S.D. dependent var	1.029	
S.E. of regression	0.663	Akaike info criterion	2.052	

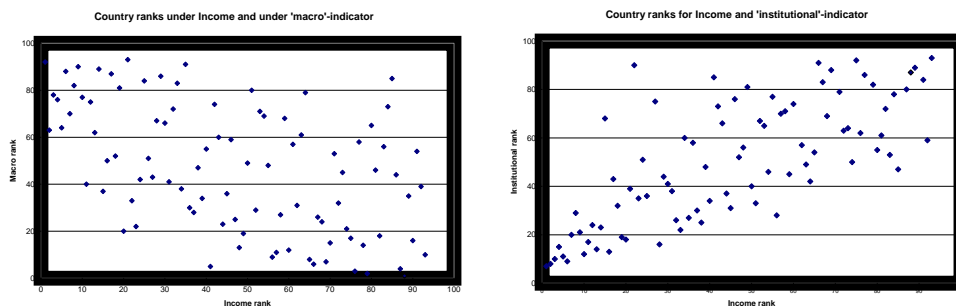


The next graphs displays scatter diagrams of the residuals from models explaining income by one of the two indicators alone, versus those from the model using both indicators together. Residuals from the model including only the "institutions"-indicator have a linear correlation of .987 with the residuals of the complete model, suggesting that this indicator does not leave unexplained much more than the two indicators together. On the other hand, the correlation coefficient between the residuals with the "macro" indicator and the residuals of the complete model is of .741, suggesting that the "institutions" adds significant explanatory power to that contained in "macro".



Nonparametric statistics can be very useful in this example, where there is no indication whatsoever that the error term in estimated models should be Normal. Spearman rank corre-

lations could be used to compare the country ranks obtained when the sample is ordered by income or according to the value of the "macro"-indicator. The rank correlation coefficient is -0.46 , with an associated t -statistic of 5.0 . A scatter plot of the ranks obtained under both orderings is shown in the left graph. Country ranks ordering by either income or "institutional"-indicator are shown to the right. The rank correlation coefficient is of 0.71 , with an associated statistic of 9.2 . Both variables contain information regarding income growth, but the information content in "institutions" seems to be greater than that in "macro".



We could now compare the country ranks obtained using income or the residuals from a projection of the "macro"-indicator on the "institutions"-indicator to analyze how much information on income growth there is in macro that it is not contained in the "institutions"-indicator. The rank correlation coefficient falls down to $-.11$ with an associated statistic of -1.0 . When we compare the income ranking with that obtained under the component of "institutions" which is not explained by "macro" the rank correlation coefficient is of $.47$, with an statistic of 4.9 . So, it also seems to be the case that most of the information in the "macro"-indicator regarding income growth is already incorporated in the "institutions"-indicator, while the latter contains information on income growth different from that contained in "macro".

5.4. Are we interpreting correctly the omitted variable bias?

A final piece of evidence comes from simple regressions from the variable to be explained on each of the explanatory variables: projecting income on "macro" reduces the standard deviation from 1.036 for the income variable to 0.890 for the residuals in that equation. Residuals from a projection of income on the "institutions" indicator alone have a standard deviation of 0.668 , versus a standard deviation of 1.029 for income in the associated subsample.

The point is that econometrics textbooks teach us to disregard these regressions as misspecified and hence, unable to produce any useful information. Unfortunately, that is not correct. Needless to say, the omitted variable bias results are right, but sometimes misinterpreted. They say that the expected value of the estimated coefficient in a regression of income on "institutions" would differ from the right one in the product of the coefficient in the omitted variable, times the coefficient obtained when projecting the omitted variable on the included one. But the latter give us precisely the change that we should expect to observe in "macro" when "institutions" change by one unit. So, the product defining the bias in the simple regression model is precisely the effect on income of the change than can be expected to arise in "macro" when "institutions" changes by one unit. That is, in expected terms, the single coefficient estimated in the simple

regression gives us the total effect on income of a unit change in "institutions". That would be the sum of the *direct effect*, produced by the unit change in "institutions", plus the *indirect effect*, that produced by the associated change we should expect to observe in "macro", because of its correlation with "institutions".

The least-squares estimator of the coefficient in a simple regression is a biased estimator of an effect on which we should hardly ever be interested, the impact on Y of a unit change in X without any change in Z . On the other hand, it is an unbiased estimator of the total impact on Y of a unit change in X , taking into account the fact that it would usually come together with a certain change in Z . This is a very different interpretation of the omitted variable bias than what we read in econometrics textbooks.

6. Time variation and tests of statistical significance

We have referred to the possibility of time variation in economic relationships as a possible source of difficulties if we mechanically apply the procedures of statistical inference. The general idea is that under time variation, we estimate an average of the true value of the parameter over the sample, and the estimation will not be very precise because we are representing the range of true values by an average.

This simple observation suggests that the loss of precision should be expected to depend on the form of the parameter variation. If a parameter oscillates over time around its average, the loss of precision will be related to the amplitude of those oscillations. If, on the other hand, the parameter has experienced a continuous increase or decrease, then substituting that behavior by an average will lead to a significant loss of precision.

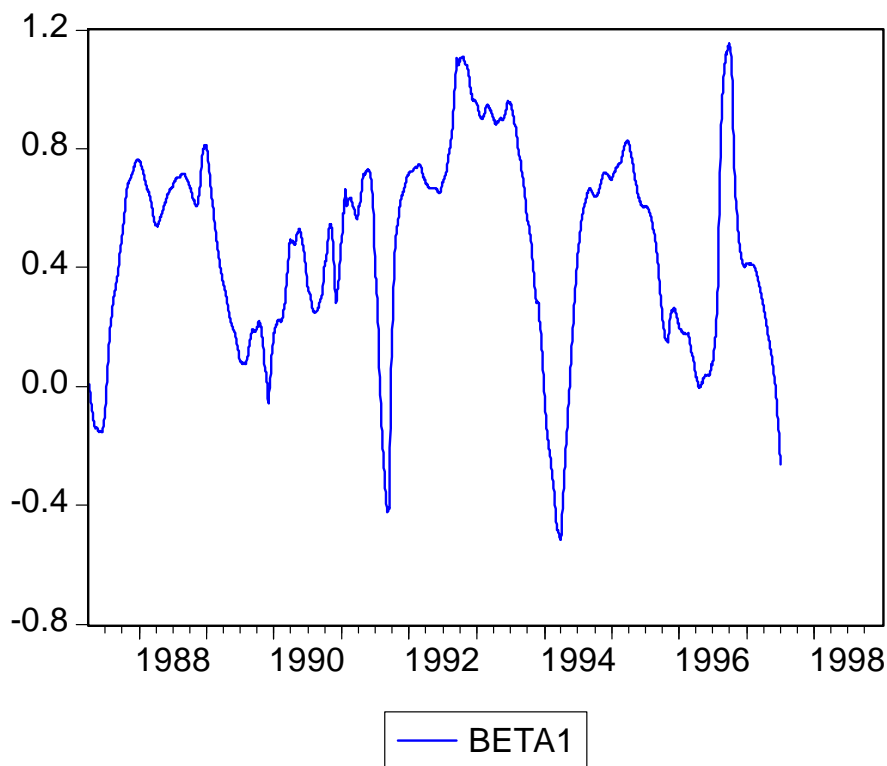
This is not a pure time series issue. In a cross-section sample, parameter dependence on a given variable will make the true, unknown value of the parameter to change with the value of that variable, leading to a loss of efficiency when estimating under parameter stability. We would in that case be talking about a nonlinear relationship, and the loss of precision that arises when that relationship is approximated by a linear model.

The loss of precision associated to parameter variation not captured by the model will bias any test in the direction of not rejecting the null hypothesis, because of the associated loss of power. The lower the precision, the wider any confidence region, so the less likely would we be to reject any null hypothesis. In particular, mechanical application of classical testing, together with the interpretation blunders we have mentioned above, would lead us to maintain too often the belief that explanatory variables do not have information content on the dependent variable.

6.0.1. An example: Testing for the Expectations Hypothesis of the term structure of interest rates

An example from the finance literature may illustrate our previous remarks. In her doctoral dissertation P. Abad tested the Expectations Hypothesis using zero-coupon interest rates obtained from *IRS* (interest rate swaps) markets in several currencies. At some point, she did recursive estimation, using each time 250 daily observations, which approximately amount to one year of market negotiation.

For the least-squares projection of the one-year zero coupon rate on the lagged one-year forward rate, recursively estimated coefficients were:



showing a significant amount of time variation. Estimates are even closer to zero than to one for most of the sample, but they experience very significant changes in magnitude. Graphs of this kind pose a serious problem, specially because a standard test performed on a single estimate of a supposedly constant coefficient do not reject the null hypothesis $H_0 : \beta = 1$. Without any more analysis, the researcher may be happy on the belief that the Expectations Hypothesis has been validated and yet, the opposite seems to be closer to be true once we have an estimation of the fluctuations on the value of β over the sample.

7. More sophisticated models

The criticism made in the previous sections goes beyond regression analysis and linear models. It is a criticism of the mechanical use that economists make too often of statistical inference tools. By the way of examples, consider the analysis of possible asymmetric effects on the volatility of the return of a given asset. Using t -tests by themselves to conclude whether these effects exist might be subject to the difficulties described in previous sections, in spite of having been obtained from efficient maximum likelihood estimation. The t -would still be comparing the numerical value of the coefficient associated to the asymmetric effect, to its standard deviation,

and the precision issue will play in this analysis the same role as described above. On the contrary, a detailed comparison of the conditional volatility series produced by the models with and without the asymmetry effect will be more appropriate. Scatter plots with the usual number of data points in high-frequency finance may not yield evidence on differences that occur at some points in time, but we can still use non-parametric statistics to discuss the similarity between the two conditional volatility series.

Something similar can be said about the literature on non-stationarity or on cointegration. Tests are applied using specific confidence levels and asymptotic critical values. A whole variety of tests have been proposed to accommodate different features of the data, like breaks, heteroskedasticity, and the like. It is often forgotten that problems with estimated models arise when the residuals are nonstationary, while a model that incorporates nonstationary variables may be used for most purposes so long as it produces stationary residuals. Sometimes, the evidence on possible nonstationarity is clear one way or the other, and the researcher knows what relationship model should then be specified. When the evidence is not so clear, we may as well proceed as if the data is stationary and carefully analyze whether residuals are stationary. Interesting contributions to the discussion about estimating in levels versus differences when stationarity is an issue are Sims, Stock and Watson, *Econometrica* (19XX)] and Sims and Uhlig (19XX), which are recommended because of being out of the mainstream in this literature.

This is also a situation in which we may want to impose the restrictions imposed by theory and analyze the results, rather than testing for the theoretical constraints. For instance, the Expectations Hypothesis would imply that nonstationary forward rates should be cointegrated with futures spot rates, with cointegrating vector (1,-1). We may test for cointegration using Johansen's approach and then test for the unit value of the cointegrating constant using the estimated standard deviations. Alternatively, we could impose the constraint, compute the spread between current short-term rates and appropriately lagged forward rates, and test for stationarity. Results will not be exactly the same. As in many other areas, we should examine the cointegration issue between these two variables from a variety of approaches. They will produce concurrent evidence which should lead the researcher to reach a final conclusion even if, as it should be expected, the suggestions received from each approach are less than fully coincident.

8. An alternative approach to testing for statistical significance

8.1. The role of the analyst of economic data

9. Steps that should be followed

9.0.1. Step 1: Data description

1. The main idea must be to avoid any methodological dogmatism. No statistical approach must be considered superior to any other in absolute terms. It is not only that the adequacy of one or other method depends on each specific situation, but also that a good research must center on a single, well defined question, and analyze it to the light of the available sample information always using a variety of methods, graphical and statistical, parametric and nonparametric.

2. The question being analyzed may require a specific type of data and sample length. The marginal propensity to consume we measure in time series data and in cross section data are different concepts. Our estimates and the results of our inference analysis may depend on the sample considered. Using samples from different units may add relevant information. Long-run properties require a long period of time, not necessarily many data points. High-frequency data may be too noisy, so more data does not amount to more information, and does not necessarily lead to increased precision.
3. Start by summarizing sample information: observed values for each variable, whether the variable has been discretized, display frequencies through an histogram, show maximum, minimum and some quantiles, mean and median, and provide always detailed information on extreme values. Compute measures of dispersion. They are needed to interpret individual coefficients.
4. Provide information on the degree of association between each potential explanatory variable and the dependent variable.
5. Provide information on the degree of colinearity between dependent variables.

9.0.2. Step 2: Estimation

1. Most characteristics on relevance of individual variables must be known by the time the model is estimated, through the descriptive analysis in previous step. Estimate models to quantify effects, although they are based on rather strict assumptions: parameter stability, invariance of remaining variables.
2. Analysis of the relevance of a given explanatory variable must use information on the variability of associated variables. To conclude on the numerical impact of a unit change requires the difficult task of dealing with colinearity. Conclusions on the relevance of one or a block of variables are always conditional on other variables included in the model. The effect of a unit change in a single variable can be measured through a simple regression. Avoid econometrics of signs and asterisks.
3. Compare residuals from constrained and unconstrained models to discuss the validity of a given hypothesis. Do not do too much formal testing.
4. Always analyze evidence on parameter instability, even in cross section data. If there is such evidence, use an appropriate estimation technique.

9.0.3. Step 3: If we test

1. Always remember relationship between precision in estimation and power in testing
2. Do not conclude on the validity of the null hypothesis when it is not rejected. Explore the test conclusion on local alternative null hypothesis.
3. Information content on a given variable is always conditional.

9.0.4. Step 4: Conclusions

1. Avoid too strong conclusions
2. Think about the role of the analyst

10. Conclusions

Statistical inference methods are often applied without the researcher being fully conscious of their scope and nature. That leads in many cases to spurious conclusions. When a more sound statistical analysis is done on the same data sets, the conclusion obtained may or may not agree with those initially proposed in the research paper. Specially alarming is the standard practice regarding hypothesis testing in general, and significance tests in particular, as well as the way how individual coefficients are usually interpreted. Power functions are never displayed, the information content of a given variable is evaluated in multivariate relationships models, and the relative importance of two explanatory variables is compared on the basis of estimated coefficients or t -statistics. Statistical significance and explanatory power are two different concepts, whose identification is unjustified, contrary to the usual rethoric among applied economists.

11. Appendix. A survey of practice in significance: applied econometrics papers in the American Economic Review, by D. Mc Closkey and S. T. Ziliak

1. Does the paper use a small number of observations, such that statistically significant differences are not found at the conventional levels merely by choosing a large number of observations?
2. Are the units and descriptive statistics for all regression variable included?
3. Are coefficients reported in elasticity form, or in some interpretable form relevant for the problem in hand and consistent with economic theory, so that readers can discern the economic impact of regressors?
4. Are the proper null hypothesis specified?
5. Are coefficients carefully interpreted?
6. Does the paper eschew reporting all t or F -statistics or standard errors, regardless of whether a significant test is appropriate?
7. Is statistical significance at the first use, commonly the scientific crescendo of the paper, the only criterion of "importance"?
8. Does the paper mention the power of the tests?
9. If the paper mentions power, does it do anything about it?
10. Does the paper eschew "asterisk econometrics"?

11. Does the paper eschew "sign econometrics"?
12. Does the paper discuss the size of the coefficients?
13. Does the paper discuss the scientific conversation within which a coefficient would be judged "large" or "small"?
14. Does the paper avoid choosing variables for inclusion solely on the basis of statistical significance?
15. After the crescendo, does the paper avoid using statistical significance as the criterion of importance?
16. Is statistical significance decisive, the conversation stopper, conveying the sense of an ending?
17. Does the paper ever use a simulation (as against a use of the regression as an input into further argument) to determine whether the coefficients are reasonable?
18. In the "conclusions" and "implications" sections, is statistical significance kept separate from economic, policy and scientific significance?
19. Does the paper avoid using the word "significance" in ambiguous ways, meaning "statistically significant" in one sentence and "large enough to matter for policy or science" in another?