Five methodological fallacies in applied econometrics

D.A. Hollanders (Technical University Delft, Netherlands)

You may post comments on this paper at
http://rwer.wordpress.com/2011/09/06/rwer-issue-57-hollanders/

Abstract
This paper discusses five methodological problems in applied econometrics. These are the problem of measurement, data mining, publication bias, the Duhem-Quine critique, and the non-repeatable nature of historical events. These problems form a third source of error next to two other more common sources of error in econometrics, sampling error and specification error. The paper argues that these problems aggravate the already difficult task of testing, but can often be dealt with. In some circumstances however testing itself is inappropriate, and econometrics is better understood as a means for description than for testing.

JEL classification: B40, C18, C50. Key words: applied econometrics, methodology, philosophy of science.

1. Introduction

Econometrics is a collection of probability statements. That is, estimated regression-coefficients come with a disclaimer that they may be wrong. The disclaimer takes the form of the probability that the estimated coefficient would have popped up if the true value of the coefficient equals zero. If this probability is below 5%, the coefficient is said to be significant. In a world of finite data the probability a true hypothesis is rejected – the type I error-never equals zero and it has to be accepted as a fact of life, even if the model is correctly specified.

A second concern, besides this Type I error, relates to the if-statement that the model is correctly specified. In its basic from, the ordinary least squares-model assumes regressors exogenous, error terms homoskedastic and uncorrelated, and sometimes normally distributed. Violations of these kind of model-assumptions form a second source of error, called specification error. Testing model-assumptions and thereby enabling correct inference is the core business of econometrics. A large part of econometrics consists then, in the words of Hendry [1980] of ‘test, test, test’.

To deal with misspecification error and being aware of sampling error distinguishes good econometrics from bad econometrics. Nonetheless exclusively focusing on these two sources of error suggests there are no other possible sources of error. There are. The paper concerns itself with five such sources of error in econometrics; these fall outside the two categories mentioned, and together constitute what may be called methodological errors. These are (1) measurement error, (2) data mining, (3) Duhem-Quine critique, (4) publication bias, (5) historical events being sui generis.

These methodological concerns are not new and no claim to novelty is made. All the same, they are treated, if mentioned at all, non-systematically in many econometric textbooks, and are dealt with, if treated at all, ad hoc in applied work. It may therefore be useful to group and categorize them. The five concerns all circle around the question
if and how it is possible to test, or if one likes, they form some epistemological disclaimers that comes with testing.

2.1. Measurement problem and the problem of conceptualizing

The saying has it that without data, you're just another person with an opinion. Without data, at least you're a person without an econometric model, as data are the sine qua non for any estimation procedure. With the data the first problems however immediately arise.

A first problem with data is error in their measurement. Individual-level data as age, occupation, gender and voting behavior may often be assumed to be measured accurately, but this is generally not the case for macro-variables as inflation, GDP, social capital, inequality and employment. Perhaps counter intuitively, measurement error in the dependent variable is a relatively minor problem, only causing the estimators to be less efficient by increasing the variance of the error term. For the independent variables, things are different. An ill-measured regressor generally introduces endogeneity, thereby causing the estimator to be biased and inconsistent, see for example Davidson and McKinnon [2004]. It is then remarkable that much effort is made (and rightfully so) to control for reversed causality, cofounders and self-selection, and that measurement error of independent variables typically is not treated with similar concern. Sometimes ill-measured instrumental variables replace an endogenous variable to control for endogeneity, inviting via the backdoor what has been tried to get rid of via the front door.

A second and related problem is that though it is often clear something is measured, it is not always equally clear what is measured exactly. Contrary to political science, not much attention is paid in economics to operationalization and conceptualization of variables. Some examples illustrate the point.

Inequality is a frequently used concept in public debates and academic disputes. It also appears in many a regression. An often used operationalization is the Gini coefficient. The Gini lies between zero (if everyone has the same income) and 1 (if one person earns all nation's income), and the higher it is, the more unequal a country is. For example, it is reported to be 0.3 for rich countries and 0.5 for Latin-American countries. The Gini is one operationalization of income inequality. There are others; for example the percentage of total wages that is earned by the 10% richest people (or some other percentage than 10%). Another possibility is to look at the part of national income that goes to the production factor labor. Yet another possibility is the mean wage divided by the median wage (the higher this value, the higher inequality). Besides the choice of the exact operationalization, there are some other bridges to cross. For one thing, income and wages were used indiscriminately in the above, while not quite being the same thing. It remains also to determine whether to look at individual or household income, whether to use pre-tax or after-tax income, and whether to include non-monetary aspects of inequality. And it makes more sense to look at life-time income of individuals than at income at a certain moment in time, If, for example, everyone earns 1 in the first period of life and 3 in the second period in life, there is no inequality in life-time income, however the Gini at each point in time would be larger than that of a (poorer) country where half of the population earns 1 all the time, and the other half 2. Last but not least, there is the problem which data to apply the exact operationalization to. One may use either tax-records or questionnaires. All in all, there is an embarrassment of riches when it comes to measuring the concept of inequality. And then, measurement error has not been mentioned yet.

One has to make a judgment call, based on the practical consideration of data-availability and the theoretical question the data have to answer. Of course, if all different operationalizations had the same qualitative result, things would not be that problematic. This is not the case however. Atkinson and Brandolini [2003] measured the same concept, the Gini, for different reliable data-sets for the Netherlands. The result was that the Gini-coefficient went up in one case, went down in another, and followed a U-shaped form in yet another. Consequently any position can be backed by a convenient choice of a data–set which in itself is reliable.

Inequality does not stand alone. GDP (Gross domestic product) is based on both historical figures and estimations and is frequently revised downward or upward afterwards. Sometimes guesswork takes the upper-hand, as the estimates of China's GDP by the Worldbank show; till recently these figures were extrapolated from a study of prices in America and China, dating back to the 1980s. Recent new price figures suggest that China's GDP may have been overestimated by 40%. This estimation applies to GDP measured at purchasing power parity, which takes into account that the yuan-equivalent a dollar has more purchasing power in China than in the US. However, something could as well be said for measuring GDP at market exchange rates, as these are the rates at which countries trade with one another. Besides measurement error and educated guess work, it is questionable what GDP is an operationalization of. If it viewed as a yardstick for economic welfare, it is first of all better to look at GDP per capita than total GDP, which is however reported in news-headlines. If this correction is applied than Japan witnessed higher economic growth than the US in 2003-2007 instead of the other way round. Correcting for population growth is only the beginning. GDP leaves out important economic factors as leisure, inequality, and the environment. Correcting for inequality makes France the richest nation (wealthier than the US), whereas correcting for leisure puts the Netherlands in first place. The US is only the winner when looking at GDP per head. If one doesn't win a single race, one may also increase the GDP by including some sectors previously not counted. Greece increased its GDP by a quarter by including inter alia smuggling, white-washing and prostitution.

Inflation can likewise be measured in different ways, for example excluding volatile prices such as food and fuel (core inflation), including them (the headline inflation) or focusing on either consumer or producer price inflation. These choices matter, or as the Economist reads 'That was the mistake made in the 1970s, when officials deluded themselves that inflation was under control by excluding ever more prices from the indices.' Nowadays, central banks focus consistently on one measure. Then it still matters which measure is looked at, or as again the Economist reads, 'the Fed focuses on "core" inflation (which excludes food and fuel) whereas the ECB targets overall inflation, America's central bank runs a looser policy in response to higher oil prices, thus pushing the dollar down.' (A different question is whether consumer good prices are that relevant when increased money supply mainly inflates asset- and house prices; only to inflate consumer price goods later with a vengeance.)

The motto of the Chicago school is that when you cannot measure, your knowledge is meager and unsatisfactory. This is an agreeable statement. However, the cases in which we cannot measure are more numerous than the cases we can. And then, like the case of inequality, the words of the political scientist Gary King [1986] hold that 'replacing the unmeasurable by the unmeaningful is not progress'.

2.2. Data mining

How does one know a theoretical plausible hypothesis holds? Test it. Does one know whether the outcome of the test is correct? One does not, but one can specify the probability with which one doesn’t. How to know the tests themselves are appropriate? That is the question of the next three paragraphs, dealing with data mining, publication bias and the Duhem-Quine critique.

Sherlock Holmes stated that 'It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.' True this may be in the circumstance of crime investigation, the principle does not apply to testing. In a crime investigation one wants to know what actually happened: who did what, when and how. Testing is somewhat different.

With testing, not only what happened is interesting, but what could have happened, and what would have happened were the circumstances to repeat itself. The particular events under study are considered draws from a larger population. It is the distribution of this population one is primarily interested in, and not so much the particular realizations of that distribution. So not the particular sequence of head and tails in coin flipping is of interest, but whether that says something about a coin being biased or not. Not (only) whether inflation and unemployment went together in the sixties is interesting, but what that tells about the true trade-off between these two economic variables. In short, one wants to test.

The tested hypothesis has to come from somewhere and to base it, like Holmes, on data is valid procedure (even more, it is good thing, or as Kennedy [2002] states 'Some economists seem to feel that data-driven theory is, somehow, unscientific. Of course, just the opposite is true.'). The theory should however not be tested on the same data they were derived from. To use significance as a selection criterion in a regression equation constitutes a violation of this principle. Sometimes for good reasons, sometimes for reasons that are not obviously convincing.

Consider for example time series econometrics. The Box-Jenkins framework explicitly models a time series as a function of its own lagged values. In doing so, inclusion of lags is based on the empirical consideration whether they are significant. Now, if for example five lags are indiscriminately tested with 5% significance one after another –which is not unusual- then the probability one lag will turn out be significant is larger than this 5%. Assuming independence between the five tests, the probability is not 5% but 22.6%. (Of course, for lags that are a multiple of a lower lag –for example 2 and 4-the tests are not independent.) While 5% is the significance level communicated, included lags are typically not really significant at the 5% level.

True, in time series there may be a good reason to capitalize on autocorrelation. It may not be clear a priori which lags matter, while it is clear that some definitely do. Theoretical ignorance then makes lag-selection an empirical question, and time series models are sometimes referred to as models of ignorance because of it. So, the Box-Jenkins framework models the auto-correlation structure of a series as good as possible first, postponing inference to the next stage. In this next stage other variables or their lagged values may be related to the time series under study. While this justifies why time series uses data mining, it leaves unaddressed the issue of the true level of significance.

In cross-section regression there is generally no such justification for data mining. Theoretically relevant regressors are easier to line up than in time series where it is clear lagged values matter but not which lag

All the same, this is sometimes recommended in a general-to-specific approach where the most general model is estimated and insignificant variables are subsequently discarded. As superfluous variables increase the variance of estimators, omitting irrelevant variables this way may increase efficiency. Problematic is that variables were included in the first place because they were thought to be (potentially) relevant. If then for example twenty variables, believed to be potentially relevant a priori, are included, then one or more will bound to be insignificant (depending on the power, which cannot be trusted to be high). Omitting relevant variables, whether they are insignificant or not, generally biases all other estimates as well due to the well-known omitted variable bias. The data are thus used both to specify the model and test the model; this is the problem of estimation. Without further notice this double use of the data is bound to be misleading if not incorrect. The tautological nature of this procedure is apparent; as significance is the selection criterion it is not very surprising selected variables are significant.

The table illustrates the point by quantifying it. Running several experiments (like testing a variable) with significance level 5% one experiment will be significant only by chance. The table gives the probability at least one test is significant as a function of the total number of experiments. With five experiments already the overall type I error is close to the earlier mentioned number of 22,6%, and with 100 it is close to one.

A practical solution is to adjust the significance level. One procedure that precisely does that is the Bon-Feroni correction, which divides the overall significance level by the number of experiments (for example, testing two independent experiments at significance level 0,025 leads to an overall significance level of 5%).

Number of independent experiments	P[at least 1 trial significant] without correction	P[at least 1 trial significant] with Bon-Feroni correction
1	0,05	0,05
2	0,10	0,049
5	0,23	0,049
10	0,31	0,049
20	0,64	0,049
100	0,994	0,049

A second solution is to double-check whether the relation holds for different sub-sets and perform out-of-sample tests. This addresses the problem, as long as these subsequent tests themselves do not become the selection criterion.

Besides running regressions with different variables, which is avoidable, it is hard to completely avoid specifying different model-specifications with the same data. More often than not, remedying heteroskedasticity, using logs of a variable instead of the variable itself or determining whether a random or fixed effect is appropriate are integral parts of the regression analysis itself, and are not decided a priori. Of course, this is exactly the point of econometrics: ‘test, test, test’ and with non-experimental data it is virtually unavoidable as running another experiment is impossible. It is however not clear how to interpret results, since properties of estimators are derived under the assumption the model is estimated only once. At least, it is then advisable to practice what Tinbergen called kitchen-sink econometrics, being explicit about all steps in the process. Next to that, adjusting the significance level comes a way to addressing the concern of data mining and fitting more than one model.

Friedman [1991] stated 'I have been extremely skeptical of relying on projections from a multiple regression, however well it performs on the body of data from which it is derived; and the more complex the regression, the more skeptical I am. (..) Regression analysis is a good tool for deriving hypotheses. But any hypothesis must be tested with data or non-quantitative evidence other than that used in deriving the regression or available when the regression was derived.'

Indeed econometrics is a good device for testing a theory that was derived from considerations unrelated to those data. Taken literary, this means that one model-specification and one only may be estimated or that the significance level be adjusted.

2.3. Duhem-Quine critique

Thus far it was tacitly assumed that it is in principle possible to test one hypothesis independently. According to the Duhem-Quine critique not so and the following makes clear what the problem is. Beck [2006] states: 'Suppose we regress Congressional vote for the incumbent on campaign spending by the incumbent. Suppose we find almost no relationship. We might conclude that money does not matter and that everyone who thought that money did matter was wrong. This would be consistent with this regression. (..) But no student of elections would stop here. Theory would tell us that challenger spending matters, and perhaps increased incumbent spending is related to increased challenger spending. Or, perhaps incumbents in trouble spend more to offset their troubles. The electoral analyst would then incorporate these theoretical ideas (...) into more appropriate regressions, which would then yield more believable results.' There is much to agree with here, in particular being critical about estimation-results.

All the same, the quoted text also raises questions. What would a student of elections do if there was still no relationship between vote for the incumbent and campaign spending by the incumbent in the extended and more appropriate model? Would the outcome of no relation be convincing, or would the student not stop then either, and if not, when, if ever, will (s)he? On another note, would the student have stopped if the relation was significant in the first model, which was however not appropriate? In other words, was the inappropriateness of the model or the insignificance of the relation the criterion to continue regressing? And if the first was the case, why not run the appropriate model from the start?

These questions refer to a more general question: if a hypothesis is rejected, is the hypothesis itself given up or is some auxiliary hypothesis rejected? Any hypothesis, like the one here, is tested under the (implicit) assumption the rest of the model is correctly specified. An (in)significant result may either lead to giving up this assumption, or to the rejection (tentative acceptation) of the tested hypothesis. In the example, it is not the hypothesis of no relation that is (tentatively) accepted, but the model in which it was tested is rejected.

This exemplifies the Duhem-Quine critique that it is not possible to test a single hypothesis; only a body of interrelated hypotheses and auxiliary-hypotheses can. This has consequences. As the whole model is tested, it is a matter of choice which one of the hypotheses making up the model is rejected (or tentatively accepted). This choice cannot be made on formal grounds. It is a matter of common sense one could say, and to a degree and in the example of Beck that is definitely the case. But it is not really the point. If for some hypothesis common sense is considered a better arbiter than a test, then there is no need to test in the first place. It also begs the question for which hypotheses common sense is a better judge than tests. May be that too can be decided by common sense, but the problem with common sense is that it is not that common, and even if so, it is not necessarily sensible. Common sense could be wrong. That is why testing is performed in the first place, to put common sense to the test.

The same holds for for example wage equations. Suppose education has a negative return, which goes against everything, from earlier econometric findings to virtually every theory, from common sense to personal experience (for some at least). Would someone rush to publish that (s)he found the revolutionary finding of a negative return? (And if so, would it be published?)

He/she probably would, but only after having made sure that every rival interpretation could be discarded. Possible rival interpretations include that (i) a control variable was omitted (ii) the specification was wrong, it should have been non-linear, (iii) the data were incorrect, (iv) heteroskedasticity was overlooked. May be education turns out to not have a negative impact on wages after all.

Or maybe it does. In that case, the result is all the more convincing. Actually, that is exactly the point. If negative returns on education are not accepted right away, positive returns shouldn't either. Otherwise some hypotheses are more equal than others. May be positive returns should be assessed even more critical. The meaning of testing is to try and falsify the received wisdom, not adhering to it.

One way to partly deal with this problem is robustness checks. It is more convincing if a hypothesis is rejected in several models that are reasonable than rejecting a hypothesis in one model as if that is the only reasonable one.

2.4. Publication bias

Private vices, economists learn, often lead to public benefits; in testing however private virtues do not necessarily add up to public benefits. That is, if only significant results are published, then results published are not significant. This also holds for tests that are performed entirely correct, and crucially hinges on whether insignificant results are published.

To take an illustrative example, every now and then, a monkey hits the newspapers for outperforming the benchmark by picking his portfolio via throwing darts at different stocks on a dart board. Considering how many highly educated and well paid people try to outperform it, a remarkable success indeed.

Though remarkable, it is not necessarily convincing. One would like to know how many monkeys are throwing darts around the world. If for example 1000 apes are doing so, and only one did the job, then the hypothesis that one could better (and cheaper) hire a monkey as portfolio manager cannot be said to be significant. One can never be sure as long as one does not know how many monkeys darted (or other species for that matter). Crucial thing here is that monkeys not outperforming the market generally do not make it into the newspaper. Considering no monkey has been employed as market-analyst, I would stick to the hypothesis it was just one of the outliers.

Not only monkeys try to outperform the markets. With some more success, academics do too. Following the Capital-Asset-Pricing-Model as the analyzing framework, asset-returns should be determined by a single factor, the (market)risk. In principal other factors should not add anything. Somehow some seem to do. A long list of effects has been put forward, inter alia, the size effect (big stocks have lower return), the momentum effect (going short in stocks that went down), and the value premium (greater return for value stocks than growth stocks). The search for effects on asset prices other than risk has been likened to handing out data-fishing licenses. One of several objections against these kind of (in itself interesting) results is indeed data-snooping-or publication bias. (Other concerns are measurement error, survivorship bias by only considering stocks of firms that did not go broke along the way, and inadequate measure of the market portfolio.) What counts is how many researchers were fitting how many effects (id est, how many darts were thrown by the researchers). Perhaps stock-returns of stocks beginning with the letter A are significantly high, but then with 26 letters (and assuming independence) one letter will be significant with a probability of 0.74. The relation should be tested on new data.

The same holds for yet another example, wage regressions. Left-handedness has been find to significantly matter for the height of wages. May be it does, but one wonders if hair color, height, weight, and eye color were all tested too by someone else.

When such tests are performed by the same researcher, this is equivalent to data mining, and the Bon-Feroni correction should have been applied. When different researchers perform different regressions with the same independent variable, this cannot be done. In that case, what holds for darting monkeys, also holds for regressing econometricians. If only significant results are published, the published results are not significant. The solution is that journals accept null-findings and insignificant results become easier to get published. In fact, after estimating a good model, one can calculate the power of the tests, id est the probability of an insignificant result. This is (approximately) the frequency with which insignificant results should be reported. If not, either something unlikely has occurred or the null hypothesis the model was correct should be rejected.

2.5. Events

Outcomes of voting polls come with standard errors and confidence intervals. Elections themselves don’t. Polls are estimates based on samples, and standard errors tell how confident one can be that the estimates would be similar were another sample drawn from the same population. On Election Day one does not need a confidence interval, as the result is known.

This points to the more general question whether the sample used in a regression is indeed a sample drawn from a larger population, or that the “sample” in fact contains the whole population. So, whether the data are equivalent to a voting poll or to Election Day; in the first case, standard errors make sense, in the latter they might not.

In some applications, such as wage equations, there is indeed a larger sample (all people working). In other applications that is less clear. Suppose one runs a panel data-regression with fixed effects for democratic countries in the 20th century, controlling for inter alia the decade (for example regressing growth and population size on size of Social Security). What one is saying then is this: France in 1910 is like England in 1990, except for it being France and not England, and except for it being 1910 and not 1990. The position is then that, after controlling for relevant variables including time and country, these two different countries in two different historical episodes can be considered draws from a larger distribution. It is rather difficult to see what that distribution is.

Three interpretations are however on offer. First, that the countries are draws from a super-or meta-population. The error-term in the regression could have turned out different. This is taking the error term too literally. The error term is in the model because we cannot explain everything in reality, not because it is really there in reality. The idea of a meta-population comes down to the belief that we live in a Panglossian world or that God plays dice. Even if one believes so, it is difficult so see this as anything else than indeed a belief.

A second interpretation is that every year another realization is drawn, which constitutes the distribution. But England in 1950 is difficult to compare to England in 1850, 1700 (or 1500), also after controlling for ‘time-effects’. Likewise it is difficult to imagine that England nowadays will be comparable, in the sense of stable regression-coefficients that is, to England in 2200. This would be equivalent to proposing that all elections in the last 100 years are draws from a larger population, and that results from elections in the twenties are generalizable to the next election. In that case one would want to know the standard errors on Election Day. Few people however do.

A third interpretation is that the sample of all democratic countries is a draw from the population of all countries, democratic or not. That is, other countries might also become democratic, and the regression tells what the regressions-coefficients are if they decide so. It is again difficult to see how a democratic country in Western-Europe is sufficiently similar to – let’s say-an African country turning democratic. And if this counter-factual is reasonable, then reversing the logic would lead to the position that one could infer what would happen to social security in a Western Europe turning despotic, by looking at African countries. At the very least, few people use the standard error in American elections to predict Dutch voting outcomes.

When regressing one takes the position that the units of analysis are similar, only differing in their values for the regression variables. And by presenting significance levels one is apparently taking the position that there is some underlying population. Even if the first assumption is acceptable –that some countries in some period are similar-the second assumption is far more problematic. The alternative is to view the countries in this period as a historical episode, and consequently the regression as a historical description. Asked what he feared most, Harold MacMillan famously replied, ‘events, my boy, events.’ Econometricians should be equally worried by them.

The unique character of historic episodes makes life in one way easier for a researcher. Statistical significance does not play an important role then; the coefficients are what they are, as one has the whole population. Regression is then a historical description of a certain episode. And if history indeed does not repeat itself, forecasting is not relevant anyway. Regression analysis does have merit, as regression can analyze historically and locally valid relations in a systematic way no other research may hope to do. In the words of Hoover: ’Econometrics is not about measuring covering laws. It is about observing unobvious regularities’ (as quoted in Kittel [2004]). These unobvious regularities are however as interesting as they are historical.

2.6. Some other concerns

In the above some methodological concerns were sketched. Before turning to the conclusion, here are two other minor concerns listed that not made it on the short list.

A first one is that statistical significance is not the same as economic importance. A variable may be significant, yet have a minor impact on the dependent variable, these are the cases the coefficient is small. It is therefore important to assess what the effect of a change of the independent variable has on the dependent variable. This is more pressing with large samples. Having many data points is of course nothing but a good thing, but it is good to bear in mind that any coefficient not literary equal to zero will flash significant if the sample size increases. And in social sciences it is hard to come up with a variable that will have absolutely nothing to do whatsoever with the dependent variable (eye color might have something to do with your wage). Nothing but a good thing, but the effect may be tiny. Small effects may be relevant (for example if wage discrimination of women is significant) or may be less relevant (for example if education raises wages by a very small amount).

A second concern is sketched by Leamer [1983]. He shows that different specifications lead to substantial different outcomes. His example is the question whether the death penalty lowers the murder rate by deterrence. The independent variable is the murder rate. Different researchers may think of different control variables, dependent on their theoretical view on what determines crime in the first place. Leamer gives five possible theoretical priors. For example a ‘right winger’ will view as crucial other deterrence variables like the probability of conviction and of execution (given being convicted). This contrasts with someone with the ‘bleeding heart’ prior, who will see economic conditions as unemployment and inequality as the prime cause of crime. Each researcher will control for the variables (s)he sees as crucial, treating all other variables as doubtful. Leamer subsequently shows that depending on whether doubtful variables are included, the ‘right winger’ may find that the drop in the murder rate per execution lies between 0.86 and 22.56. This is in itself already a large range, but the ‘bleeding heart’ may find the effect to lie between a drop of 25.6 and an increase in the murder rate of 12.37. Leamer states the feeling ‘that any inference from these data about the deterrent effect of capital punishment is too fragile to be believed’. This indicates that a priori theories, on which the model is and should be based, matters for the final outcomes. Different theories lead to different outcomes and outcomes may just end up confirming the theories on which they were based in the first place.

3. Conclusion

The above has raised some points related to testing. These methodological concerns are to be distinguished from the usual problem of econometrics such as heteroskedasticity, auto-correlation, non-stationarity, and endogeneity.

The first point, measurement error and error of measurement, indicates that even before testing begins, the problems already started. Especially measurement error of independent variables can have severe consequences, as it introduces endogeneity. A related problem is that entities as inflation, inequality and GDP can be operationalized in different ways, leading to truly different outcomes. It is remarkable that very precise estimates are based on data that are not very precise.

Three other concerns -publication bias, Duhem-Quine method and data mining-lie at the core of testing itself. These are difficult to avoid completely, but can reasonably be dealt with. When multiple tests are performed significance levels should be adjusted accordingly. Robustness checks should be performed and researchers should be explicit about the ways they have tried to falsify their result and which steps have been taken in the estimation procedure. Finally, it would be good that null-findings are just as easy to publish as significant results.

It may be useful to realize that these problems are also present in natural sciences. There too measures may come with error, insignificant results may be hard to publish, a single hypothesis cannot be tested, and there too models are derived from the data. There is one difference, and that is a big difference: the possibility of running another experiment. In economics this is difficult to do and econometrics is the art of making the most of non-experimental data or as Orcutt has it ‘Doing econometrics is like trying to learn the laws of electricity by playing the radio’.

All the same, if these laws are really claimed to be general -in the sense that there is a stable relation and a wider population-then it should in principal be possible to play another radio, that is to perform another test on data that the researcher was not and could not have been aware of when estimating the model. Testing the model on these new data is the real test, as the words of Friedman also suggests.

Critics of experimental economics doubt its relevance for its lack of external validity. A valid point, but if the external validity of econometrics is much better, it should be possible to conduct another test. There are circumstances in which such a test is not only difficult to imagine but simply impossible to perform. It is not possible to run the 20th century over again to see whether macro-economic relations in democratic countries still hold. The fifth concern articulates this. It is for example not obvious that there is a population of democratic countries in the 20th century out of which countries were drawn. Then the vocabulary of testing is just not appropriate. As history does not repeat itself, it is also not relevant either, as forecasting is not important anyway. Regression analysis of for example OECD-countries should then be understood as a descriptive account first and foremost in which testing is impossible or irrelevant, or both. Coefficients cannot be said to be significant or not, they are what they are.

There remains a valuable role to play for econometric models; they are able to describe historical events in a systematic way no qualitative researcher can hope to do. And it can, by doing so, suggest historical relationships that are not visible by other means.

Hicks stated that 'as economics pushes on beyond "statics", it becomes less like science, and more like history'. Similarly, as econometrics pushes beyond repeatable events, it becomes more like history, though hopefully not less like science. In those cases econometrics is a collection of historical statements.

References

Atkinson, A. B. and A. Brandolini [2003], 'Promise and Pitfalls in the Use of "Secondary Data-Sets: Income Inequality in OECD Countries as a Case Study', *Journal of Economic Literature*, 39(3).

Beck, N. [2006], 'Is Causal-Process Observation an Oxymoron?', *Symposium on Rethinking Social Inquiry*, 347-352.

Davidson, R. and McKinnon, J.G. [2004], *Econometric Theory and Methods*.

Ericsson, N. R, [2004], 'The ET interview: professor David F. Hendry', *International Finance Discussion Papers*.

Friedman, M. and A. J. Schwartz [1991], 'Alternative Approaches to Analyzing Economic Data', *American Economic Review*, 81(1), 39-49.

Hendry, D. F. [2002], 'Applied Econometrics without sinning', *Journal of Economic Surveys*, 16(4), 591-603.

Hendry, D. F. [1980], 'Econometrics-Alchemy or Science?', *Economica,* 47(188), 387-406.

Kennedy, P. E. [2002], 'Sinning in the Basement: what are the rules? The ten commandments of applied econometrics', *Journal of Economic Surveys*, 16(4), 569-589.

Keuzenkamp, H. A. and J. R. Magnus [1995], 'On tests and significance in econometrics', *Journal of Econometrics*, 67, 5-24.

King, G. [1986], 'How not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science', *American Journal of Political Science*, 30(3), 666-687.

Kittel, B. [2004], 'Een gekke methodologie? Over de grenzen van macrokwantitatief sociaal-wetenschappelijk onderzoek', *Vossiuspers UvA*.

Leamer, E. E. [1983], 'Let's take the Con Out of Econometrics', *American Economic Review*, 73(1), 31-43.

The Economist,
'A less fiery dragon?', December 1st 2007, pp. 82.
'Grossly Distorted picture', February 11th 2006, pp. 70.
'Grossly Distorted picture', March 15th 2008, pp. 92.
'Feeling the heat', June 24th 2006, pp. 81-82.
'Ben's bind', May 3rd 2008, pp. 79-80.

'En wéér klopt Jacko de AEX-index', *de Telegraaf*, December 28th 2008.

'Economie Griekenland gered door prostituees', *De Volkskrant*, September 30th 2006, pp.7.

Author contact: d.a.hollanders@uvt.nl

SUGGESTED CITATION:
D.A. Hollanders, "Five methodological fallacies in applied econometrics", real-world economics review, issue no. 57, 6 September 2011, pp. 115-126, http://www.paecon.net/PAEReview/issue57/Hollanders57.pdf

You may post and read comments on this paper at
http://rwer.wordpress.com/2011/09/06/rwer-issue-57-hollanders