An Introductory-Level Explanation of my Critique of AT99

Ross McKitrick

August 25, 2021

The AT99 article "Checking for model consistency in optimal fingerprinting" is here.
My critique of this article is here.
My first blog post about it (at Judith Curry's site) is here.

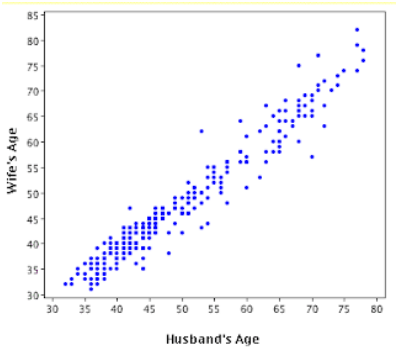
1 INTRODUCTION

My article in *Climate Dynamics* shows that the AT99 method is theoretically flawed and gives unreliable results. A careful statement of the implications must note an elementary principle of logic. Remember that, according to logic, we can say "Suppose A implies B; then if A is true therefore B is true." Example: all dogs have fur; a beagle is a dog; therefore a beagle has fur. But we cannot say "Suppose A implies B; A is not true therefore B is not true." Example: all dogs have fur; a cat is not a dog, therefore a cat does not have fur. But we can say "Suppose A implies B; A is not true therefore we do not know if B is true." Example: all dogs have fur; a dolphin is not a dog, therefore we do not know if a dolphin has fur.

In this example "A" is the statistical argument in AT99 which they invoked to prove "B"—the claim that their model yields unbiased and valid results. I showed that "A", their statistical argument, is not true. So we have no basis to say that their model yields unbiased and valid results. In my article I go further and explain why there are reasons to believe the results will typically be invalid. I also list the conditions needed to prove their claims of validity. I don't think it can be done, for reasons stated in the paper, but I leave open the possibility. Absent such proof, applications of their method over the past 20 years leave us uninformed about the influence of GHG's on the climate. Here I will try to explain the main elements of the statistical argument.

2 REGRESSION

Most people are familiar with the idea of drawing a line of best fit through a scatter of data. This is called linear regression. Consider a sample of data showing, for example, wife's age plotted against the husband's age.



Clearly the two are correlated: older men have older wives and vice versa. You can easily picture drawing a straight line of best fit.

The formula for a straight line is $Y = a + bX$. Here, Y and X are the names of the variables. In the above example Y stands for wife's age and X stands for husband's age. a and b are the *coefficients* to be estimated. b is the slope coefficient. When you draw the line of best fit you are selecting numerical values for a and b. We may be interested in knowing whether b is positive, which implies that an increase in X is associated with an increase in Y. In the above example it clearly is: any reasonable line through the sample would slope upwards. But in other cases it is not so obvious. For example:

Here a line of best fit would be nearly horizontal, but might slope up. For the purpose of picturing why statistical theory becomes important for interpreting regression analysis it is better to have in mind the above graph rather than the earlier one. We rarely have data sets where the relationship is

as obvious as it is in the husband-wife example. We are more often trying to get subtle patterns out of much noisier data.

It can be particularly difficult to tell if slope lines are positive if we are working in multiple dimensions: for instance if we are fitting a line $Y = a + bX + cW + dZ$ through a data set that also has variables W and Z and their coefficients c and d to contend with. Regardless of the model we need some way of testing if the true value of b is definitely positive or not. That requires a bit more theory.

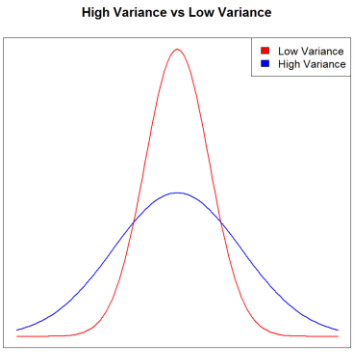
Note that regression models can establish correlation, but correlation is not causation. Older men do not cause their wives to be older; it is just that people who marry tend to be of the same age group. If we found deaths by drowning to be correlated with ice cream consumption, it would not prove that eating ice cream causes drowning. It is more likely that both occur in warm weather, so the onset of summer causes both events to rise at the same time. Regression models can help support interpretations of causality if there are other grounds for making such a connection, but it must be done very cautiously and only after rigorously checking testing whether the model has omitted important explanatory variables.

3 SAMPLING AND VARIANCE

The first example above is a plot of a *sample* of data. It is clearly not the entire collection of husbands and wives in the world. A sample is a subset of a *population*. When we do statistical analysis we have to take account of the fact that we are working with a sample rather than the entire population (in principle, the larger the sample, the more representative it is for the entire population). The line of best fit through the sample can only ever yield an estimate of the true value of b. In conventional notation we denote an estimate of b with a 'hat', writing it $\hat{b}$. Because it is an estimate, we can only really talk about a range of possible values. Regression yields a distribution of possible estimates, some more likely than others. If you fit a line through data using a simple program like Excel it might only report the central slope estimate $\hat{b}$ but what the underlying theory yields is a distribution of possible values.

Most people are familiar with the idea of a 'bell curve' which summarizes data, like the distribution of grades in a class, where many values are clustered around the mean and the number of observations diminishes as you go further away from the mean. The wideness of a distribution is

summarized by a number called the variance. If the variance is low the distribution is narrow and if it is high the distribution is wide:



Regression analysis yields an estimate both of $\hat{b}$ and its variance $v(\hat{b})$. A closely related concept is the standard error of $\hat{b}$ which is the square root of $v(\hat{b})$ and can be denoted with a Greek sigma: $\hat{\sigma}$. Statistical theory tells us that, as long as the regression model satisfies a certain set of conditions, there is a 95% probability that the true (population) value of b is inside an interval bounded approximately by $2\hat{\sigma}$ above and below $\hat{b}$. This is called the 95% Confidence Interval.

Given a sample of data on (in this case) X and Y, we can use regression methods to fit a line $Y = \hat{a} + \hat{b}X$ and if we are confident $\hat{b}$ is above zero it implies that an increase in X leads to an increase in Y. "Confident" here means that $\hat{b}$ is more than $2\hat{\sigma}$ greater than zero. If it isn't we say that the coefficient is positive but not *statistically significant*.

4 Bias, Efficiency and Consistency

The value of $\hat{b}$ is obtained using a formula that takes in the sample data and pops out a number. There are many formulas that can be used. The most popular one is called Ordinary Least Squares or OLS. It is derived by supposing that the straight line allows us to predict the value of Y that corresponds with each value of X, but there will be an error in each such prediction, and we should choose the values of $\hat{a}$ and $\hat{b}$ that minimize the sum of the squared errors. OLS also yields an estimate of the variances of each coefficient.

Expected value is a concept in statistics that refers to a probability-weighted average of a random variable. The expected value of a random variable g is denoted $E(g)$. OLS yields a distribution for $\hat{b}$, which means it has an expected value. Statistical theory can be used to show that, as long as the regression model satisfies a certain set of conditions, $E(\hat{b}) = b$. In other words, the expected value is the true value. In this case we say the estimator is *unbiased*. It is also the case that the variance estimate is unbiased (again as long as the regression model satisfies a certain set of conditions).

Since there are many possible estimation formulas besides OLS, we need to think about why we would prefer OLS to the others. One reason is that, among all the options that yield unbiased estimates, OLS yields the smallest variance.[1] So it makes the best use of the available data and gives us the smallest 95% Confidence Interval. We call this *efficiency*.

Some formulas give us estimated slope coefficients or variances that are biased when the sample size is small, but as the sample size gets larger the bias disappears and the variance goes to zero, so the distribution collapses onto the true value. This is called *consistency*. An *inconsistent* estimator has the undesirable property that as we get more and more data we have no assurance that our coefficient estimates get closer to the truth, even if they look like they are getting more precise because the variance is shrinking but does not go to zero.

5 THE GAUSS-MARKOV CONDITIONS AND SPECIFICATION TESTING.

I have several times referred to "a certain set of conditions" that a regression model needs to satisfy in order for OLS to yield unbiased, efficient and consistent estimates. These conditions are listed in any introductory econometrics textbook and they are called the Gauss-Markov (GM) conditions. Much of the field of econometrics (which is a branch of statistics focused on using regression analysis to build economic models) is focused on testing for failures of the GM conditions and proposing remedies when failures are detected.

Some failures of the GM conditions imply that $\hat{b}$ will still be unbiased, but its variance estimate is biased. So we might get a decent estimate of the slope coefficient but our judgment of whether it is

[1] This assumes we are only considering linear estimators, which is a detail we can ignore for the present purpose.

significant or not will be unreliable. Other failures of the GM conditions imply that both $\hat{b}$ and $\hat{\sigma}$ are biased. In this case the analysis may be spurious and totally meaningless.

As an example of a bad research design, suppose we have data from hundreds of US cities over many years showing both the annual number of crimes in the city and the number of police officers on the streets, and we regress the annual number of crimes on the annual number of police officers to test if crime goes down when more police are deployed. There are several problems that would likely lead to multiple GM conditions failing. First, the sample consists of small and large cities together, so the range and dispersion of the data over the sample will vary, which can cause biased variance estimates. Second, there will be lag effects where a change in policing might lead to a change in crime only after a certain amount of time has passed, which can bias the coefficient and variance estimates. Third, while crime may depend on policing, policing levels may also depend on the amount of crime, so both variables are determined by each other: one is not clearly determined outside the model. This can severely bias the coefficients and lead to spurious conclusions (such as that more policing leads to higher crime levels). Finally, both crime and policing depend on factors not included in the model, and unless those outside factors are uncorrelated with the level of policing the coefficient and variance estimates will be biased.

It is therefore critical to test for failures of the GM conditions. There is a huge literature in econometrics on this topic, which is called *specification testing*. Students who learn regression analysis learn specification testing all the way along. If a regression model is used for economics research, the results would never be taken at face value without at least some elementary specification tests being reported.

There is a class of data transformations that can be used to remedy violations of some GM conditions, and when they are applied we then say we are using *Generalized Least Squares* or GLS. Having applied a GLS transformation doesn't mean we can assume the GM conditions automatically hold, they still have to be tested. In some cases a GLS transformation is still not enough and other modifications to the model are needed to achieve unbiasedness and consistency.

6 THE AT99 METHOD

Various authors prior to AT99 had proposed comparing observed climate measures to analogues simulated in climate models with and without GHG's (which are called "response patterns") to try to

determine if including the effect of GHG's significantly helps explain the observations, which would then support making an attribution of cause. They refer to their method as "fingerprinting" or "optimal fingerprinting." Those authors had also argued that the analysis would need to be aided by rescaling the data according to local climatic variability: put more weight on areas where the climate is inherently more stable and less weight on areas where it is "noisier". To do that required having an estimate of something called the "climate noise covariance matrix" or C_N which measures the variability of the climate in each location and, for each pair of locations, how their climate conditions correlate with each other. Rather than using observed data to compute C_N climatologists have long preferred to use climate models. While there were reasons for this choice, it created many problems (which I discuss in my paper). Once C_N is obtained from a climate model, to compute the required regression weights one needs to do a bit of linear algebra: first compute the inverse of C_N and then compute the matrix root of the inverse. This would yield a weighting matrix P that would help "extract" information more efficiently from the data set.

One problem the scientists ran into, however, is that climate models don't have enough resolution to identify all the elements of the C_N matrix independently. In mathematical terms we say it is "rank deficient", and an implication is that the inverse of C_N does not exist. So the scientists chose to use an approximation called a "pseudo-inverse" to compute the needed weights. This created further problems.

7 THE AT99 ERROR

AT99 noted that applying a weighting scheme makes the fingerprinting model like a GLS regression. And, they argued, a GLS model satisfies the GM conditions. Therefore the results of this method will be unbiased and efficient. That slightly oversimplifies their argument, but not by much. And the main error is obvious. You can't know if a model satisfies the GM conditions unless you test for specific violations. AT99 didn't even state the GM conditions correctly, much less propose any tests for violations.

In fact they derailed the whole idea of specification testing by arguing that one only needs to test that the climate model noise covariance estimates are "reliable" (their term—which they did not define), and they proposed a test statistic which they called the "Residual Consistency" or RC test for that purpose. They didn't offer any proof that the RC test does what they claimed it does. For example it has nothing to do with showing that the residuals are consistent estimates of the unknown error

terms. In fact they didn't even state precisely *what* it tests, they only said that if the formula they propose pops out a small number, the fingerprinting regression is valid. In my paper I explained that there can easily be cases where the RC test would yield a small number even in models that are known to be misspecified and unreliable.

And that, with only one slight modification, has been the method used by the climate science profession for 20 years. A large body of literature is based on this flawed methodology. No one noticed the errors in the AT99 discussion of the GM conditions, no one minded the absence of any derivation of the RC test, and none of the hundreds of applications of the AT99 method were subject to conventional specification testing. So we have no basis for accepting any claims that the results of the optimal fingerprinting literature are unbiased or consistent. In fact, as I argued in my paper, the AT99 method as set out in their paper *automatically* fails at least one GM condition and likely more. So the results have to be assumed to be unreliable.

The slight modification came in 2003 when Myles Allen and a different coauthor, Peter Stott, proposed shifting from GLS to another estimator called Total Least Squares or TLS.[2] It still involves using an estimate of C_N to rescale the data, but the slope coefficients are selected using a different formula. Their rationale for TLS was that the climate model-generated variables in the fingerprinting regression are themselves pretty 'noisy' and this can cause GLS to yield coefficient estimates that are biased downwards. This is true, but econometricians deal with this problem using a technique called Instrumental Variables or IV. We don't use TLS (in fact almost no one outside of climatology uses it) because, among other things, if the regression model is misspecified, TLS over-corrects and imparts an upward bias to the results. It is also extremely inefficient compared to OLS. IV models can be shown to be consistent and unbiased. TLS models can't, unless the researcher makes some restrictive assumptions about the variances in the data set which themselves can't be tested; in other words, unless the modeler "assumes the problem away."

8 IMPLICATIONS AND NEXT STEPS

The AT99 method fails the GM conditions. As a result, its usage (including the TLS variant) yields results which might by chance be right, but in general will be biased and inconsistent and therefore

[2] Allen, M.R. and P.A. Stott (2003) Estimating Signal Amplitudes in Optimal Finger-Printing, Part I: Theory. *Climate Dynamics* 21:477—491. DOI 10.1007/s00382-003-0313-9

cannot be assumed to be reliable. Nothing in the method itself (including use of the RC test) allows scientists to claim more than that.

The AT99 framework has another important limitation which renders it unsuitable for testing the hypothesis that greenhouse gases cause changes in the climate. The method depends on the assumption that the model which generates the C_N matrix and the response patterns is a true representation of the climate system. Such data cannot be the basis of a test that contradicts the assumed structure of the climate model. The reason has to do with how hypotheses are tested. Going back to the earlier example of estimating $\hat{b}$ and its distribution, statistical theory allows us to construct a test score (which I'll call t) using the data and the output of the regression analysis which will have a known distribution if the true value of b is zero. If the computed value of t lies way out in the tails of such a distribution then it is likely not consistent with the hypothesis that $b = 0$. In other words, hypothesis testing says "If the true value of b is zero, then the statistic t will be close to the middle of its distribution. If it is not close to the middle, b is probably not zero."

For this to work requires us to be able to derive the distribution of the test statistic under the hypothesis that the true value of b is zero. In the fingerprinting regression framework suppose b represents the measure of the influence of GHG's on the climate. The optimal fingerprinting method obliges us to use data generated by climate models to estimate both b and its variance. But climate models are built under the assumption that GHG's have a large positive effect on the climate, or $b > 0$. So we can't use that data to estimate the distribution of a statistic under the assumption that $b = 0$. Such a test will be spurious and unreliable.

My work on this topic continues. I have written another paper examining the biases in TLS fingerprinting regressions. And I am working on a paper that shows the effects of applying basic specification testing to fingerprinting regressions and remedying the resulting failures.

9 REPLYING TO RESPONSES

A number of commentators on my paper have tried to shrug my criticism off as unimportant or irrelevant. But AT99 has been used hundreds of times in the climate literature and studies applying it have been cited thousands of times. Also the IPCC chose to focus on AT99 as soon as it appeared, promoting it in the 2001 IPCC Third Assessment Report (TAR Chapter 12, Box 12.1, Section 12.4.3

and Appendix 12.1), and it has been referenced in every IPCC Assessment Report since. TAR Appendix 12.1 was headlined "Optimal Detection is Regression" and began

> The detection technique that has been used in most "optimal detection" studies performed to date has several equivalent representations (Hegerl and North, 1997; Zwiers, 1999). It has recently been recognised that it can be cast as a multiple regression problem with respect to generalised least squares (Allen and Tett, 1999; see also Hasselmann, 1993, 1997)

Their reliance on AT99 continues today: see AR6 Section 3.2.1. The relevance of my critique is proven by the heavy reliance the climate profession has placed on AT99 over the years, including the nearly exclusive reliance on the RC test and the absence of any mention of the conditional independence assumption.

More specifically, in considering any response to my paper, it will be important to note whether it actually disagrees with my paper, or simply tries to change the subject. I anticipate that a lot of respondents will implicitly concede that my paper is correct, but argue it doesn't matter because so much time has gone by. However, as a matter of the scientific record it is important to understand and acknowledge if AT99 made errors in their mathematical presentation and whether the subsequent literature corrected those errors or simply carried them forward. As far as I have seen, they were carried forward in the sense that people still to this day rely on the RC test and they still use AT99-type regression models without testing for specification errors associated with specific GM conditions.

Also, and more generally, if major errors in AT99 went unnoticed for so long, it calls into question how much confidence we can have in the various other methodologies that have been developed in climate journals in subsequent years. Having worked on paleoclimate reconstruction methods, trend estimation and comparisons methods, and now on optimal fingerprinting, comparing climate journals to stats or econometrics journals I find that climate journals seem to rely on referees who don't know how to ask the right questions when confronted with a novel statistical method. My discussion of the introduction of the RC test contrasts what AT99 did with what you'd expect to see in a statistics or econometrics journal.

Another line of response has been that AT99 has been superseded by regularization methods (associated with Ribes, Terray, Hannart and so forth) and they get the same results. I mention this approach in my paper in a couple of places. Regularization is an alternative to the Moore-Penrose algorithm to estimate the inverse of the non-invertible climate noise matrix. It yields a full-rank approximation so there is no longer a dependence on the rank truncation parameter K. But there remains the problem of showing that the resulting estimator is consistent (see condition [N3] in my paper). It's a computational improvement, possibly, but not a theoretical one. The regularization literature has never discussed the conditional independence assumption nor has it revisited the claims around the Gauss-Markov theorem applying.

There are some other recent attribution methods, including time series methods (such as cointegrating vector autoregressions or CVAR) that do not make any use of climate models. My critique does not specifically apply to these. There may be other issues with them, but I haven't looked at them in detail. The ones I have seen have largely been confined to analysing the time series of global average surface temperatures and have considered only a very limited number of explanatory variables.