Methodology in Practice: Statistical Misspecification Testing

Deborah G. Mayo and Aris Spanos†

The growing availability of computer power and statistical software has greatly increased the ease with which practitioners apply statistical methods, but this has not been accompanied by attention to checking the assumptions on which these methods are based. At the same time, disagreements about inferences based on statistical research frequently revolve around whether the assumptions are actually met in the studies available, e.g., in psychology, ecology, biology, risk assessment. Philosophical scrutiny can help disentangle 'practical' problems of model validation, and conversely, a methodology of statistical model validation can shed light on a number of issues of interest to philosophers of science.

1. Introduction. Methodological disputes that arise in practice often turn on questions of the nature, interpretation, and justification of methods and models that are relied on to learn from incomplete, and often 'observational' (or nonexperimental), data: the methodology of statistical inference and statistical modeling. Philosophers of science who immerse themselves in practice have often discovered that the methods and models routinely used to these ends are mired in confusion, controversy, and unquestioned assumptions, often along with a reluctance among practitioners to tinker with already accepted methods. Although at a number of junctures these methodological debates revolve around philosophical and logical issues, the scientific community seldom looks to philosophy for help in their resolution.

This symposium (to which our paper is a contribution) arose from our thinking that this situation should be remedied. Although precisely how to make progress here is open to debate, at least one point of agreement that emerged from our discussions (with Clark Glymour, Kristin Shrader-

†To contact the authors, please write to: Deborah G. Mayo, Department of Philosophy, 235 Major Williams Hall, Virginia Tech, Blacksburg, VA 24061; e-mail: mayod@vt.edu. Aris Spanos, Department of Economics, 3016 Pamplin Hall, Virginia Tech, Blacksburg, VA 24061; e-mail: aris@vt.edu.

Philosophy of Science, 71 (December 2004) pp. 1007–1025. 0031-8248/2004/7105-0032$10.00

Frechette, and Alison Wiley) is that it will require serious collaborative efforts between philosophers of science and scientists, and that establishing and promoting such collaboration demands a much more serious commitment to interdisciplinary work than is now standard. Such a 'practical-philosophical' enterprise, we think, is very much of a two-way street: On the one hand, philosophers of science bring relevant skills and perspectives to bear on methodological problems in practice, not because of a uniquely privileged stance, but because of a penchant for broad generality, logical criticism, and articulating reasoning; on the other, such collaborative work seems to hold significant promise for a more dynamic, effective, and interesting philosophy of science.

As much as these general points deserve analysis in their own right, rather than pursue them here we propose to jump right into illustrating a portion of our collaborative efforts concerning model validation in the social sciences. One set of concerns in statistical modeling has to do with gaps between variables in a statistical model and primary factors or questions of interest, but an even more basic question is whether the assumptions needed to reliably model the statistical variables are met, e.g., whether an assumption of independent trials is marred by dependencies. Our focus here is the latter, although the two are interrelated. We will thus be talking about a methodology for testing misspecifications in statistical models, misspecification (M-S) testing.

A full methodology of M-S testing, as we see it, would tell us how to specify and validate statistical models, and how to proceed when statistical assumptions are violated. So developing such a methodology requires methods for uncovering and probing model assumptions, isolating sources of any anomalous results, and iterative procedures for accommodating any flawed assumptions in respecified models until arriving at a *statistically adequate* model—a model that is adequate for subsequent (primary) statistical inferences.

It is important to note at the outset that the problem of statistical model specification is distinct from model selection, insofar as the latter selects from an *assumed family of models* according to one or another criterion, e.g., Akaike (AIC), Bayesian Information Criterion (BIC), by N-P testing, and by causal structure searches. M-S testing is thus of central importance to all model selection approaches because the presence of misspecifications jeopardizes the ground for model selection, whichever criterion one uses. Progress in the area of M-S tests would thus benefit the considerable, interesting literature among philosophers of science revolving around these model selection techniques (e.g., Forster 2001; Forster and Sober 1994; Glymour 1997; Spirtes, Glymour, and Scheines 2001; Woodward 1988).

Our example will focus on the linear regression model (LRM), which

forms the backbone of most statistical models of interest, as it gives rise to variants such as non-Normal, nonlinear, and/or heteroskedastic regression models, as well as multivariate and structural equations models. A central problem we consider is that of spurious regression or *spurious correlation*. Despite the extensive literature on spurious correlation, going back to the late nineteenth-century Yule (1895), the source of the problem and the way it intertwines with the problem of misspecification remain ill understood. Traditional tools that are routinely used, we argue, are poor at detecting the most obvious sorts of violations in assumptions; they indicate a strong relationship between variables x_t and y_t when in fact the two variables are unrelated, and they endorse fallacious ways of accommodating violations when they are found. Our aim, as we proceed, is to zero in on the most philosophically interesting questions and problems.

2. Problems of Validation in the Linear Regression Model (LRM). The *Linear Regression Model* (LRM) may be seen to take the form

$$M_0\colon\ y_t = \beta_0 + \beta_1 x_t + u_t, \quad t = 1, 2, \ldots, n, \ldots,$$

where $\mu_t = \beta_0 + \beta_1 x_t$ is the systematic component, and $u_t = y_t - \beta_0 - \beta_1 x_t$ is the error (nonsystematic) component. The error process $\{u_t, t = 1, 2, \ldots, n, \ldots\}$ is assumed to be Normal, Independent, and Identically Distributed (NIID) with mean 0, variance σ^2, i.e., Normal white noise. Using the data $\mathbf{z} := \{(x_t, y_t), t = 1, 2, \ldots, n\}$ the coefficients (β_0, β_1) are estimated (by least squares) yielding an empirical equation intended to enable us to understand how y_t varies with x_t.

2.1. Empirical Example. In his attempt to find a way to understand and predict changes in the U.S. population, imagine that an economist discovers, using regression, an empirical relationship that appears to provide almost a 'lawlike' fit:

$$\begin{aligned} y_t &= 167.115 + 1.907x_t + \hat{u}_t, \\ &\quad (.610)\ \ (0.024) \end{aligned} \tag{1}$$

where y_t denotes the population of the U.S.A. (in millions), and x_t denotes a secret variable whose identity we will reveal later on. Both series refer to annual data for the period 1955–1989, and the numbers in brackets denote standard errors for the coefficient estimates.

A primary statistical question. A primary question under the LRM is: *How good a predictor is* x_t? The goodness of fit measure of this estimated regression, $R^2 = .995$, indicates an almost perfect fit. Testing the statistical significance of the coefficients (whether they differ from 0) shows them

to be highly significant: p-values are nearly 0, indicating a very strong relationship between the variables.[1] The question now is: *Is this inference reliable?* We can answer this affirmatively only if data $\mathbf{z}$ satisfy the probabilistic assumptions of the LRM, i.e., the errors are NIID with mean 0, variance σ^2.

Misspecification (M-S) tests: 'secondary' questions. Questions of model validation may be tackled by M-S tests, which can be regarded as 'secondary' questions in relation to the primary statistical ones. Whereas primary statistical inferences take place *within* a specified (or assumed) model M, the secondary inference has to put M's assumptions to the test; so to test M's assumptions, we stand *outside* M, as it were.

Strictly speaking, however, the hypotheses of interest in M-S tests are

$$H_0\text{: the assumption(s) of statistical model } M \text{ hold for data } \mathbf{z},$$

as against alternative not-H_0, where not-H_0 would consist of all of the ways one or more of M's assumptions can fail. However, this alternative is too unwieldy; in practice one needs to consider a specific departure from H_0, i.e., a specific way in which H_0 can be false, in order to apply a statistical significance test to H_0. The logic of such significance tests is this: We identify a test statistic $d(\mathbf{Z})$ to measure the distance between what is observed $\mathbf{z}_0$ and what is expected assuming the null hypothesis H_0, so as to derive the distribution of $d(\mathbf{Z})$ under the assumption of H_0. If $\mathbf{z}_0$ is improbably far from what is expected under H_0, i.e., if

$$P(d(\mathbf{Z}) > d(\mathbf{z}_0);\ H_0 \text{ true}) = p$$

is very small, then H_0 is rejected, and there is said to be evidence that the assumption(s) are violated ("p" denotes the p-value associated with the observed difference).

As soon as a distance function $d(\cdot)$ is chosen, it is important to see, one is choosing, in effect, a specific direction of departure from H_0 to be probed. Since only the null is explicitly set out, these departures may be regarded as *implicit alternatives* to the null. They need to be made explicit by considering the particular violations from H_0 that the given test is capable of probing.

2.2. Testing Randomness: A Nonparametric 'Runs' Test. M-S tests may be carried out in two ways: *nonparametric* and *parametric* tests. We begin with a nonparametric test called the *runs test*. It tests for both the in-

1. For example, in testing $\beta_1 = 0$ vs. $\beta_1 \neq 0$ our estimated β_1 is 79.5 standard deviations away from 0. Under the assumption of model M_0, this test statistic $T = \sqrt{n}\hat{\beta}_1/\sqrt{\text{Var}(\hat{\beta}_1)}$ has a known distribution (Student's t), so we can calculate the *statistical significance* of our estimate: $P(T > 79.5;\ H_0) = 0.000$.

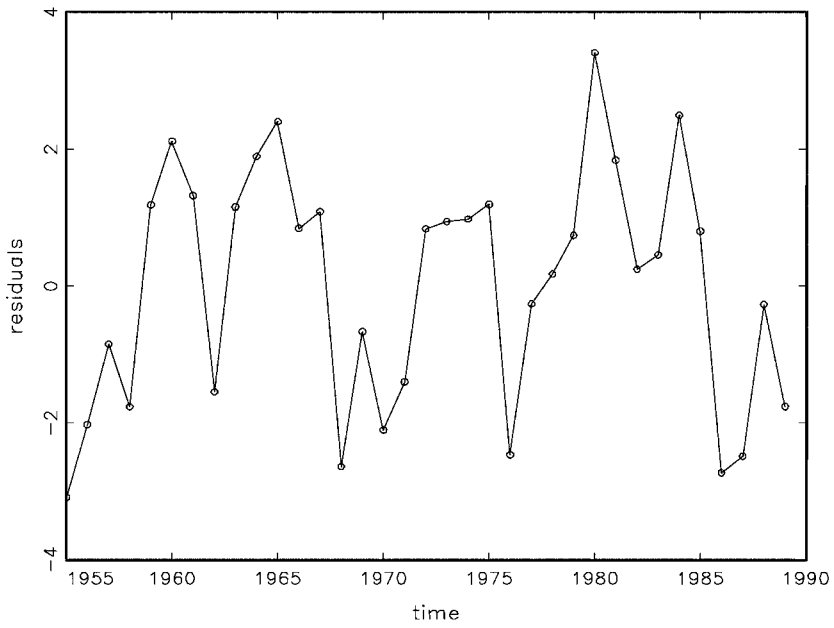

Figure 1. *t*-plot of residuals.

dependence and identical distribution (IID) assumptions at the same time, that is, it is a test for *randomness*. The randomness assumption is generally expressed as: the error u_t is IID. To address this, the runs test looks at the *residuals*,

$$\{\hat{u}_t = y_t - \hat{\beta}_0 - \hat{\beta}_1 x_t, \quad t = 1, 2, \ldots, n\},$$

from our estimated regression to see if they exhibit randomness; ($\hat{\beta}_0$, $\hat{\beta}_1$) denote the estimates of the coefficients (β_0, β_1). As is typical in M-S tests, the data set **z** (for the primary statistical test) is remodeled to ask this secondary question. Instead of the particular value of each observed residual (one for each data point) one records if the difference between successive observations is positive, a "+", or negative, a "−". The residuals from estimated regression (1), shown in Figure 1, give rise to the following pattern of ups and downs:

++−++−−+++−+−+−+++++−++++−−++−−++−

The patterns are called *runs*: a sub-sequence of one type (pluses only or minuses only) immediately preceded and succeeded by an element of the other type.

https://doi.org/10.1086/425064 Published online by Cambridge University Press

The appeal of such a nonparametric test is that its own validity does not depend on the assumptions of the (primary) model under scrutiny: we can calculate the probability of different numbers of runs just from the hypothesis that the assumption of randomness holds. In particular, were the data to have arisen from a random process, then both too many and too few runs would be very rare—indicating trends or cycles in the data—where the rareness is given by the statistical significance level or p-value corresponding to the values of the statistic R, the number of runs.[2]

As is plausible, then, the test based on R takes the form: Reject H_0 iff the observed R differs from the expected R (under IID) by a sufficient amount (in either direction), where the expected number of runs, assuming randomness, is $(2n - 1)/3$ or in our case of 35 values, 23. The data from our example yields 18 runs (around 2.4 standard deviation units)—yielding a p-value of around .02.[3] Equivalently, 98% of the time we would expect an R closer to 23, were the null hypothesis true, that is,

$$P(\text{a smaller departure from IID; IID true}) = .98.$$

So the data, we might say, are a good indication of nonrandomness.[4]

However, since the test is sensitive to departures from both the I and ID assumptions (both of which make up 'randomness'), rejecting the null does not warrant inferring anything more than a denial of IID. The test itself does not indicate whether the fault lies with one or the other or both assumptions. Given that no specific alternative is contemplated, there is no temptation to infer anything beyond 'non-IID'. Let us now compare this to a parametric M-S test.

2.3. Testing Non-autocorrelation: The Parametric Durbin-Watson (D-W) Test. The most widely used parametric test for independence is the Durbin-Watson (D-W) test. All the assumptions of the LRM are retained, except the one under test—independence—which is, as is often said, 'relaxed'. In particular, the original error term in M_0 is extended to allow for the possibility that the errors u_t are correlated with their past, i.e.,

2. Thus we can test the hypotheses about randomness by testing the following null and alternative hypotheses about the expected number of runs, $E(R)$:

$$H_0\colon E(R) = (2n - 1)/3, \quad H_1\colon E(R) \neq (2n - 1)/3.$$

3. $Z_R = (R - E(R))/\sqrt{\text{Var}(R)}$ is approximately Normally distributed ($N(0, 1)$); see Levene 1952.

4. We could say the data indicate nonrandomness with *severity* .98 (see Mayo and Spanos 2000, Mayo 1996).

$u_t = \rho u_{t-1} + \varepsilon_t$. That is, a new model, M_1, the *Autocorrelation-Corrected* (A-C) LRM, is assumed:

$$M_1\colon\ y_t = \beta_0 + \beta_1 x_t + u_t,\ u_t = \rho u_{t-1} + \varepsilon_t, \quad t = 1,\ 2,\ \ldots,\ n,\ \ldots.$$

The D-W test assesses whether or not $\rho = 0$ in model M_1. That is, it tests the conjunctions

$$H_0\colon\ \{M_1\ \&\ \rho = 0\},\ \text{vs.}\ H_1\colon\ \{M_1\ \&\ \rho \neq 0\}.$$

Applying this test to the data in our example, the D-W test statistic rejects the null hypothesis (at level .02), which is taken as grounds to adopt H_1. This move to infer H_1, however, is warranted only if we *are* within M_1. Granted, if $\rho = 0$, we are back to the LRM, but $\rho \neq 0$ does not entail the particular violation of independence asserted in H_1. Nevertheless, modelers routinely go on to infer H_1 upon rejecting H_0, despite warnings, e.g., "A simple message for autocorrelation correctors: Don't" (Mizon 1995).

A clear elucidation of the flawed reasoning—the sort of thing a philosophical analysis might afford—can highlight the fallacy that prevents this strategy from uncovering what is really wrong both with the original LRM and the A-C LRM. However, far from detecting the fallacy, the traditional strategy finds strong evidence that the error-autocorrelation of the new model is necessitated by the data.

Consider how 'autocorrelation correctors' traditionally proceed: having inferred the A-C LRM, the next step is to estimate the new model yielding

$$M_1\colon\ y_t = \underset{(.939)}{167.209} + \underset{(0.037)}{1.898}x_t + \hat{u},\ u_t = \underset{(0.152)}{.431}u_{t-1} + \hat{\varepsilon}_t. \tag{2}$$

Has the A-C LRM 'corrected for' the anomalous result that led to rejecting the LRM? It appears that it has, at least according to the traditional analysis. The common strategy here would be to check if the new error process $\{\varepsilon_t, t = 1, 2, \ldots, n, \ldots\}$ is free of any autocorrelation (by running another D-W test), and indeed it is.[5] Whereas the A-C LRM has, in one sense, 'corrected for' the presence of autocorrelation, because the assumptions of model M_1 have been retained in H_1, this check had no chance to uncover the various other forms of dependence that could have been responsible for $\rho \neq 0$. Duhemian problems loom large. By focusing exclusively on the error term the traditional viewpoint overlooks the ways the systematic component of M_1 may be misspecified and fails also to

5. The Durbin-Watson test statistic, D−W = 1.831, is not significant. The t-test for H_0: $\rho = 0$ vs. H_1: $\rho \neq 0$, is significant ($p = .004$), indicating that the 'correction' is justified. In addition, the A-C LRM shows improvements over the LRM in fit ($R^2 = 0.996$).

acknowledge other hidden assumptions, e.g., $\boldsymbol{\theta} := (\beta_0, \beta_1, \sigma^2)$ are not changing with the index $t = 1, 2, \ldots, n$.

We are only highlighting the central logical flaws here; a full philosophical scrutiny of the traditional strategy would go further. As a result of these logical failings, the traditional strategy leads to models which, while acceptable according to its own self-scrutiny, are in fact inadequate: using them for the 'primary' statistical inferences yields actual error probabilities much higher than the ones it is thought to license, and they are unreliable when used to predict values beyond the observed data. This fallacy illustrates the kind of pejorative use of the data, to construct (ad hoc) a model to account for an anomaly, that leads many philosophers of science, as well as statistical modelers, to be skeptical of the 'double counting' of data that goes on in M-S testing; see Spanos 2000. Without pinpointing precisely where and when such double counting leads to unreliable results, however, it is impossible to identify those strategies for using data to detect and correct for violated assumptions that sidestep these difficulties.[6] Progress in identifying reliable M-S procedures, therefore, will at the same time give clues to solving the analogous problem in philosophy of science.

3. Partitioning the Space of Possible Models: Probabilistic Reduction. We would like a procedure that correctly identifies the flaws in conjectured statistical models and that also points the way to developing an adequate model. Let us explore a procedure that has been developed in order to put the entire process of model validation on a sounder philosophical footing. Recall that the task in validating a model M_0 (LRM) is actually to test 'M_0 is valid' against everything else. In other words, if we let H_0 assert that the sample $\mathbf{Z}$ follows a given distribution $f(\mathbf{z})$, the alternative H_1 would be the entire complement of M_0, more formally

$$H_0\colon f(\mathbf{z}) \in M_0 \text{ vs. } H_1\colon f(\mathbf{z}) \in [\mathcal{P} - M_0],$$

where $\mathcal{P}$ denotes the set of all possible statistical models that could have given rise to $\mathbf{z}_0 := \{(x_t, y_t), t = 1, 2, \ldots, n\}$. The traditional analysis of the LRM has already, implicitly, reduced the space of models that could be considered. It reflects just one way of reducing the set of all possible models of which data $\mathbf{z}_0$ can be seen to be a realization. This provides the motivation for the modeling approach developed by Spanos (1986, 1989, 1995).

6. The issue of when and why the 'use-novelty' requirement has an epistemological rationale is an old one (Mayo 1991, 1996).

TABLE 1. THE LINEAR REGRESSION MODEL (LRM)

$y_t = \beta_0 + \beta_1 x_t + u_t,\ t = 1, 2, \ldots, n, \ldots$		
[1] Normality:	$D(y_t \mid \mathbf{x}_t;\ \theta)$	Normal
[2] Linearity:	$E(y_t \mid X_t = x_t) = \beta_0 + \beta_1 x_t$	Linear in x_t
[3] Homoskedasticity:	$\mathrm{Var}(y_t \mid X_t = x_t) = \sigma^2$	Free of x_t
[4] Independence:	$(y_t \mid X_t = x_t),\ t \in T$	Independent
[5] t-homogeneity:	$\boldsymbol{\theta} := (\beta_0, \beta_1, \sigma^2)$	t-unvarying

Given that each statistical model arises from the joint distribution,

$$D(\mathbf{Z}_1, \ldots, \mathbf{Z}_n;\ \boldsymbol{\phi}) := D((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n);\ \boldsymbol{\phi}),$$

we can consider how one or another set of probabilistic assumptions on the joint distribution gives rise to different models like the LRM. The assumptions come from a menu of three broad categories: (D) Distribution, (M) Dependence, (H) Heterogeneity. For example, the LRM arises when we reduce $\mathcal{P}$ by means of the assumptions (called *reduction assumptions*),

(D) Normal (N), (M) Independent (I), and (H) Identically Distributed (ID).

Since we are partitioning or reducing $\mathcal{P}$ by means of the probabilistic assumptions, it may be called the Probabilistic Partitioning or *Probabilistic Reduction* (PR) approach.[7] The same assumptions traditionally given by means of the error term are specified in terms of the observable random variables (y_t, x_t): [1]–[4] (see table 1). This has several advantages, especially if one is attempting to get at the foundations. Hidden or implicit assumptions now become explicit ([5]). Moreover, the LRM (conditional) assumptions can be assessed *indirectly* from the data via the (unconditional) reduction assumptions, since N entails [1]–[3], I entails [4], and ID entails [5].

As a first step, we partition the set of all possible models coarsely in terms of reduction assumptions on $D(\mathbf{Z}_1, \ldots, \mathbf{Z}_n;\ \boldsymbol{\phi})$ as follows:

7. This is because when the NIID assumptions are imposed on $D(\mathbf{Z}_1, \ldots, \mathbf{Z}_n;\ \phi)$ the latter simplifies into a product of conditional distributions $D(y_t \mid \mathbf{x}_t;\ \varphi_1)$ (LRM):

$$D(\mathbf{Z}_1, \ldots, \mathbf{Z}_n;\ \phi) \overset{\mathrm{I}}{=} \prod_{t=1}^{n} D_t(\mathbf{Z}_t;\ \varphi_t) \overset{\mathrm{IID}}{=} \prod_{t=1}^{n} D(\mathbf{Z}_t;\ \varphi) = \prod_{t=1}^{n} D(y_t \mid \mathbf{x}_t;\ \varphi_1) D(\mathbf{x}_t;\ \varphi_2).$$

https://doi.org/10.1086/425064 Published online by Cambridge University Press

	LRM	Alternatives (Coarsely Partitioned)
(D) Distribution	Normal	Non-Normal
(M) Dependence	Independent	Non-I
(H) Heterogeneity	Identically Distributed	Non-ID

Given the practical impossibility of probing for violations in all possible directions, the PR approach consciously considers an effective probing strategy to decide on the directions in which the primary statistical model might be potentially misspecified. Having taken us back to the joint distribution, why not get ideas by looking at y_t and x_t themselves? This is what the PR approach prescribes. Although such graphical techniques involve a kind of 'double use' of data (or 'use-constructed hypotheses' (Mayo, 1991, 1996)), far from being a pejorative form of data-snooping (as some allege), they become a powerful way to get ideas about which M-S tests to apply in order to assess violations of the reduction assumptions (and indirectly the model assumptions) most effectively and most severely.

3.1. Learning from Graphical Techniques: t-Plots. Plotting the observed data—y_t, population of the USA in millions, x_t, secret variable—over time (1955–1989) we get time plots or *t-plots*; Spanos (1999), as in Figures 2 and 3.

We ask: What would be expected if each data series were to have come from a NIID process, as is assumed by the LRM? We answer this by simulation, giving the graph in Figure 4. When we compare this typical realization of a NIID process with the *t*-plots of the two series given in Figures 2–3, we can see that the data exhibit glaring departures from IID. In particular, both data series show the mean is increasing with time—i.e., strong mean-heterogeneity (trending mean). The traditional approach described above did not detect the presence of mean-heterogeneity and so it misidentified the source of the problem with the LRM.

We can summarize our progress in finding a useful alternative to probe thus far:

	LRM	Alternative (to Probe)
(D) Distribution	Normal	?
(M) Dependence	Independent	?
(H) Heterogeneity	Identically Distributed	Mean-heterogeneity

3.2. Discriminating and Amplifying the Effects of Mistakes. We could correctly assess dependence if our data were ID and not obscured by the

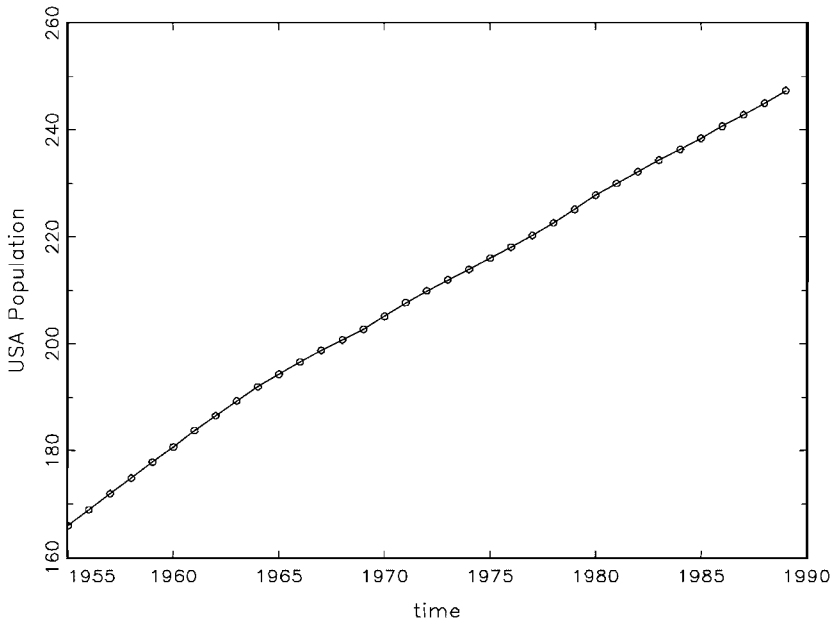

Figure 2. U.S. population (y_t).

influence of the trending mean. Although we can not literally manipulate relevant factors, we can 'subtract out' the trending mean in order to learn what it would be like if there were no trending mean, by "manipulations on paper" (Mayo 1996) yielding detrended x_t and y_t.

The data in both Figures 5 and 6 exhibit, to a trained eye, positive dependence or 'memory' in the form of cycles—Markov dependence. So the independence assumption also looks problematic, explaining the autocorrelation detected by the M-S tests discussed earlier. Our LRM assessment so far, just on the basis of the graphical analysis, is:

	LRM	Alternative (to Probe)
(D) Distribution	Normal	?
(M) Dependence	Independent	Markov
(H) Heterogeneity	Identically Distributed	Mean-heterogeneity

Finally, we could evaluate the distribution assumption (Normality) graphically if we had IID data, so if we could see what the data $\{(x_t, y_t),\ t = 1, 2, \ldots, n\}$ would look like without the heterogeneity ('detrended') and without the dependence ('dememorized'), we could get some ideas about the appropriateness of the Normality assumption.

https://doi.org/10.1086/425064 Published online by Cambridge University Press

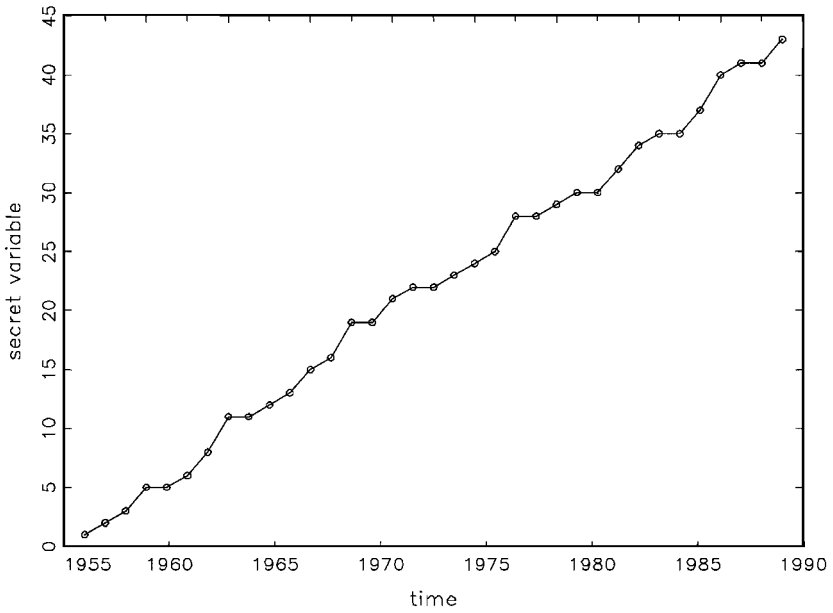

Figure 3. Secret variable (x_t).

These plots in Figures 7 and 8, as well as the scatter-plot of (x_t, y_t), show no telltale signs of departure from Normality (the scatter-plot is elliptical). So our procedure has brought us to a specific type of alternative to the LRM:

	LRM	Alternative (to Probe)
(D) Distribution	Normal	Normal
(M) Dependence	Independent	Markov
(H) Heterogeneity	Identically Distributed	Mean-heterogeneity

While there are still several selections under each of the headings of Markov dependence and mean-heterogeneity, the length of the Markov dependence (m), and the degree (l) of the polynomial in t, would be discerned in subsequent rounds of the probing strategy. The model derived by imposing this set of reduction assumptions on $D(\mathbf{Z}_1, \ldots, \mathbf{Z}_n; \boldsymbol{\phi})$ is the

https://doi.org/10.1086/425064 Published online by Cambridge University Press

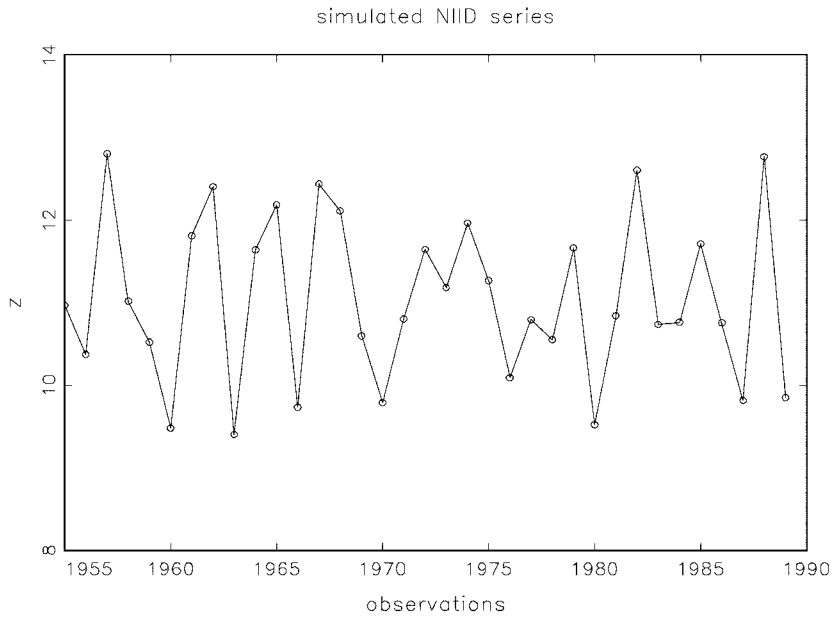

Figure 4. A typical realization of a NIID process.

Dynamic Linear Regression Model (DLRM) (l, m), $l > 0$, $m > 0$):

$$y_t = \beta_0 + \beta_1 x_t + \overbrace{\sum_{i=1}^{l} \delta_i t^i}^{\text{trending mean}} + \overbrace{\sum_{j=1}^{m} \alpha_j y_{t-j} + \gamma_j x_{t-j}}^{\text{temporal dependence}} + u_t,$$

$$t = 1, 2, \ldots, n, \ldots.$$

The values of (l, m) chosen on statistical adequacy grounds[8] are: $l = 1$, $m = 2$. The DLRM is arrived at by probatively 'looking at the data' through the graphical discernments, but we must be clear on what is licensed by such qualitative assessments.

3.3. The Nature of the Inferences from Graphical Techniques. What is the status of the learning from graphs? As we see it, the graphs enable one to get good ideas about the kinds of violations for which it would

8.

$$y_t = \underset{(5.122)}{17.687} + \underset{(0.080)}{0.193}t - \underset{(.036)}{.000}x_t + \underset{(0.147)}{1.496}y_{t-1} + \underset{(.037)}{.013}x_{t-1} - \underset{(0.148)}{0.560}y_{t-2} + \underset{(.035)}{.014}x_{t-2} + \hat{u}_t,$$

$R^2 = 0.9999$, $s = 0.154$, $n = 35$.

https://doi.org/10.1086/425064 Published online by Cambridge University Press

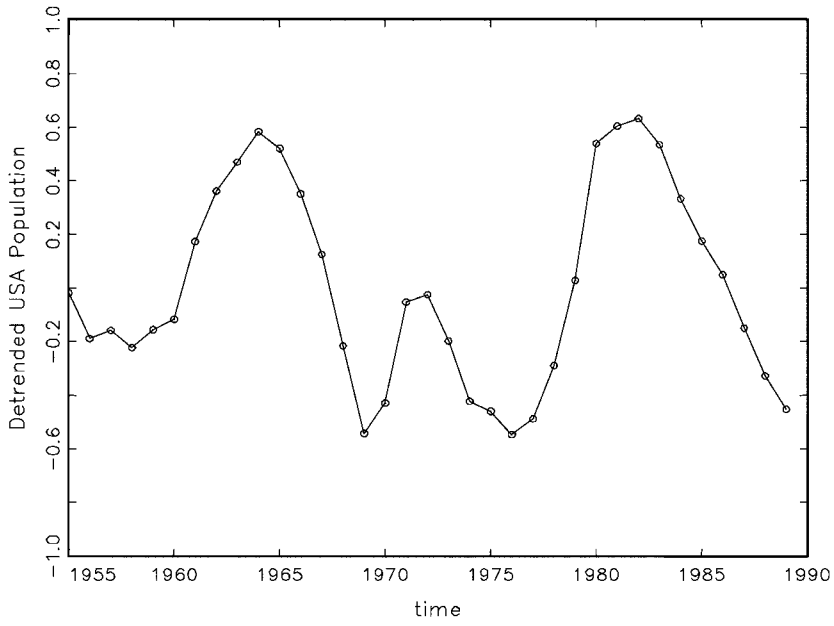

Figure 5. Detrended population (y_t).

be useful to probe, much as looking at a forensic clue (e.g., footprint, tire track) helps to narrow down the search for a given suspect, or a fault-tree, for a given cause.

Consider the first t-plots looking for trends, Figures 2–3. Without posing a formal question, one reasons that such a trending t-plot would only have arisen were the mean of the underlying process to be changing systematically with t. The same discernment can be achieved with a formal analysis (using the nonparametric runs test), perhaps more discriminating than can be accomplished by even the most trained eye, but the reasoning and the justification are much the same. If there is a license to infer evidence of nonrandomness with the runs test, then so would there be with the informal graphical analysis. All of this, of course, invites further philosophical examination—both logical and empirical.[9]

The combined indications from the graphs constitute evidence of departures from the LRM in the direction of the DLRM, but only, for the moment, as a fruitful model to probe further. We are not licensed to infer it is itself a statistically adequate model until its own assumptions are

9. The empirical component involves simulating data deliberately generated to violate or obey the various assumptions.

https://doi.org/10.1086/425064 Published online by Cambridge University Press

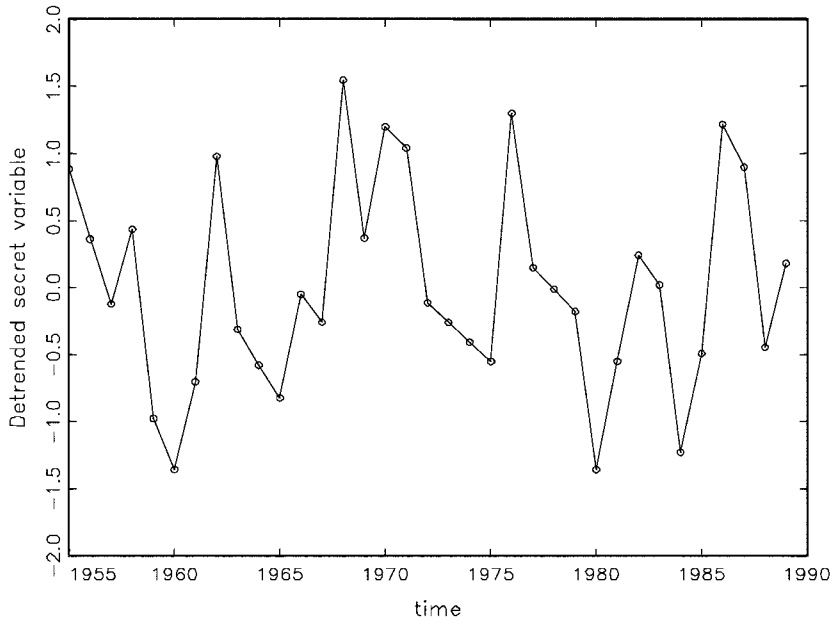

Figure 6. Detrended secret variable (x_t).

subsequently tested. Even when they are checked and found to hold up—which happens to be in this case—our inference must still be qualified. While we may infer that the model is statistically adequate, this should be understood only as licensing the use of the model as a reliable tool for primary statistical inferences, but not necessarily as representing the substantive phenomenon being modeled.

4. Back to the Primary Statistical Inference: Nonsense Regressions. Having established the statistical adequacy of the estimated DLRM, we are then licensed in making 'primary' statistical inferences about the values of parameters in this model. In particular, we can proceed to assess the ability of the secret variable to help predict the population of the USA. A test (an F test) of joint significance of the coefficients of (x_t, x_{t-1}, x_{t-2}) does not reject the hypothesis that they are all 0, indicating that the secret variable is uncorrelated with the population variable![10] We are thus led

10. $F(3, 26) = .302[.823]$.

https://doi.org/10.1086/425064 Published online by Cambridge University Press

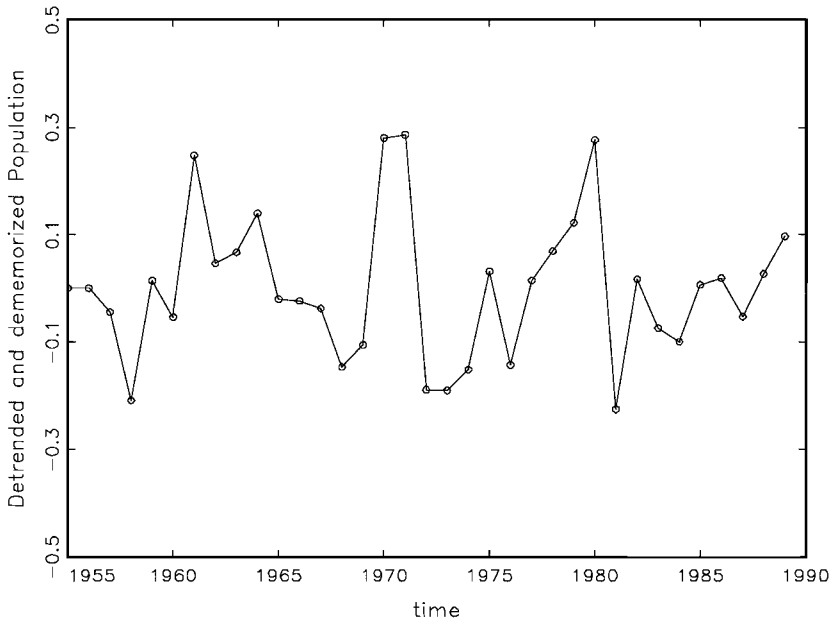

Figure 7. Detrended and dememorized population (y_t).

to drop these terms from the DLRM, giving rise to an *Autoregressive of Order m* (AR(m)) model with trends:

$$y_t = \beta_0 + \overbrace{\sum_{i=1}^{l} \delta_i t^i}^{\text{trending mean}} + \overbrace{\sum_{j=1}^{m} \alpha_j y_{t-j}}^{\text{temporal dependence}} + u_t,$$

$$t = 1, 2, \ldots, n, \ldots.$$

The estimated form of this AR(m) model yields

$$\begin{aligned} y_t = {} & 17.148 + 0.217t + 1.475y_{t-1} - 0.572y_{t-2} + \hat{u}_t, \\ & (4.781)\ (0.063)\ \ (0.134)\ \ \ (0.119) \end{aligned} \tag{3}$$

$R^2 = 0.9999$, $s = 0.147$, $n = 35$. Hence, on the basis of a statistically adequate model we were able to infer reliably that the secret variable contributed nothing towards predicting or explaining the population variable. The regression between x_t and y_t suggested by the estimated models M_0 and M_1 turn out to be nonsense. The source of the problem is that the inferences concerning the significance of x_t were unreliable due to the fact that the underlying models were misspecified.

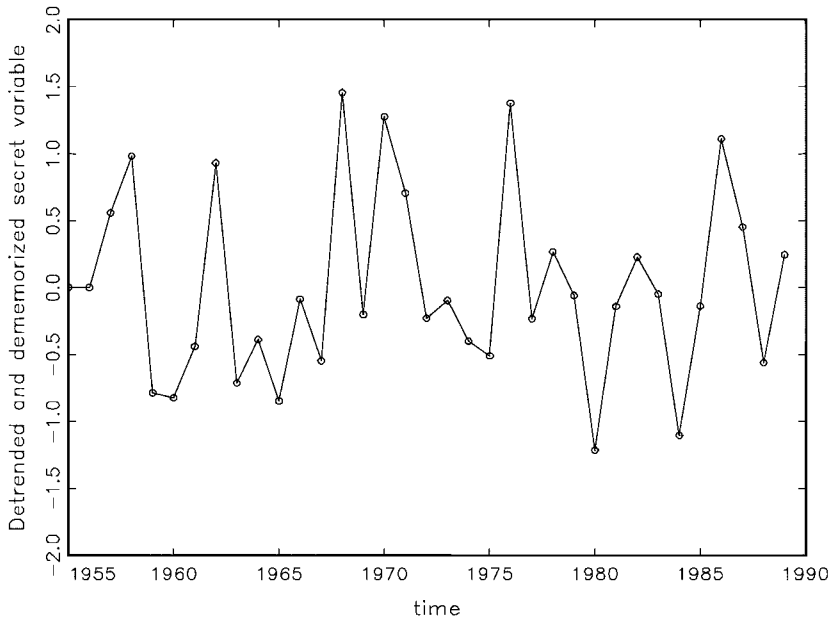

Figure 8. Detrended and dememorized secret variable (x_t).

Revealing the identity of the secret variable shows the egregiousness of such erroneous inferences. It turns out that:

x_t—the number of pairs of shoes owned by Spanos's grandmother!

While this was an extreme case, the usual methods, we saw, do not readily pick up the problem. It serves as a 'canonical exemplar' for the kind of errors for which one requires methods to probe and rule out, and for the fallacies which must conscientiously be avoided. Examples of unreliable inferences abound in the empirical literature of numerous fields, especially in the social sciences.

5. Concluding Remarks: Further Questions for a Methodology of Model Validation. This brief foray into the methodology of M-S testing raises numerous questions and problems of philosophical interest. We list three areas for further work:

1. *Justifying assumptions of tests of assumptions.* A central concern in basing statistical inference on models whose own assumptions require testing is the threat of ending in an infinite regress of testing. How can we get around this in M-S testing? The secret lies in shrewd testing strategies: following a logical order of nonparametric and parametric tests,

https://doi.org/10.1086/425064 Published online by Cambridge University Press

and combining tests that jointly probe several violations with deliberately varied assumptions. It will be important to spell this out more fully in future work.

2. *The "double counting" charge*. M-S tests "snoop" at the data, and use them to arrive at a hypothesis, e.g., the trials are not independent, as well as to evaluate that hypothesis. Such "double counting", some charge, leads to tests with high or incalculable error probabilities. Much more needs to be said to explain when this criticism has weight and when this criticism misses its mark. Far from increasing error rates, multiple tests, if appropriate, may serve to cross-validate and fortify other tests, so that the model inferred as statistically adequate has passed a reliable test.

3. *Relevance for existing model selection techniques*. Two related questions may be advanced using the PR approach: (i) How serious are the consequences of misspecification for one or another model selection technique, e.g., AIC, BIC, Neyman-Pearson, causal modeling? (ii) What would be the upshot of applying model selection techniques to the models outputted by the PR procedure? Model selection techniques require starting with a prespecified family of models in which it is assumed the true model lies. The iterative PR procedure, by contrast, with its numerous choices under the three menu items, may give rise to models that would not have arisen by the traditional model specification procedures.

The problem of whether, and if so how, to validate statistical models is of fundamental importance across the entire modeling landscape, both for practitioners wishing to ensure that the sophisticated modeling techniques that are increasingly available lead to models with predictive and explanatory power, and for philosophers seeking to justify induction and statistical inference. We hope to encourage further work in this area.

REFERENCES

Forster, Malcolm R. (2001), "The New Science of Simplicity", in Arnold Zellner, Hugo A. Keuzenkamp, and Michael McAleer (eds.), *Simplicity, Inference and Modelling*. Cambridge: Cambridge University Press, 83–119.

Forster, Malcolm R., and Elliott Sober (1994), "How to Tell When Simpler, More Unified, or Less ad Hoc Theories Will Provide More Accurate Predictions", *The British Journal for the Philosophy of Science* 45: 1-35.

Glymour, Clark (1997), "A Review of Recent Work on the Foundations of Causal Inference", in McKim and Turner (1997).

Levene, Howard (1952), "On the Power Function of Tests of Randomness Based on Runs Up and Down", *Annals of Mathematical Statistics* 23: 34–56.

Mayo, Deborah G. (1991), "Novel Evidence and Severe Tests", *Philosophy of Science* 58: 523–552.

——— (1996), *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Mayo, Deborah G., and Aris Spanos (2000), "A Post-data Interpretation of Neyman-Pearson Methods Based on a Conception of Severe Testing", *Measurements in Physics*

and Economics Discussion Paper Series, History and Methodology of Economics Group. London School of Economics and Political Science, Tymes Court, London.
McKim, Vaughn R., and Stephen P. Turner (eds.) (1997), *Causality in Crisis? Statistical Methods and the Search for Causal Knowledge in the Social Sciences*. Notre Dame, IN: University of Notre Dame Press.
Mizon, Graham E. (1995), "A Simple Message for Autocorrelation Correctors: Don't", *Journal of Econometrics* 69: 267–288.
Spanos, Aris (1986), *Statistical Foundations of Econometric Modelling*. Cambridge: Cambridge University Press.
——— (1989), "On Re-reading Haavelmo: A Retrospective View of Econometric Modeling", *Econometric Theory* 5: 405–429.
——— (1995), "On Theory Testing in Econometrics: Modeling with Nonexperimental Data", *Journal of Econometrics* 67: 189–226.
——— (1999), *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*. Cambridge: Cambridge University Press.
——— (2000), "Revisiting Data Mining: 'Hunting' with or without a License", *The Journal of Economic Methodology* 7: 231–264.
Spirtes, Peter, Clark Glymour, and Richard Scheines (2001), *Causation, Prediction, and Search*, 2nd ed. Cambridge, MA: MIT Press.
Woodward, James (1988), "Understanding Regression", in Arthur Fine and Jarrett Leplin (eds.), *PSA 88*, vol. 1. East Lansing, MI: Philosophy of Science Association, 255–269.
Yule, Udny (1895), "On the Correlation of Total Pauperism and the Proportion of Out Relief", *Economic Journal* 5: 603–611.