Tour IV More Auditing: Objectivity and Model Checking

4.8 All Models Are False

> . . . it does not seem helpful just to say that all models are wrong. The very word model implies simplification and idealization. . . . The construction of idealized representations that capture important stable aspects of such systems is, however, a vital part of general scientific analysis. (Cox 1995, p. 456)

A popular slogan in statistics and elsewhere is “all models are false!” Is this true? What can it mean to attribute a truth value to a model? Clearly what is meant involves some assertion or hypothesis about the model – that it correctly or incorrectly represents some phenomenon in some respect or to some degree. Such assertions clearly can be true. As Cox observes, “the very word model implies simplification and idealization.” To declare, “all models are false” by dint of their being idealizations or approximations, is to stick us with one of those “all flesh is grass” trivializations (Section 4.1). So understood, it follows that all statistical models are false, but we have learned nothing about how statistical models may be used to infer true claims about problems of interest. Since the severe tester’s goal in using approximate statistical models is largely to learn where they break down, their strict falsity is a given. Yet it does make her wonder why anyone would want to place a probability assignment on their truth, unless it was 0? Today’s tour continues our journey into solving the problem of induction (Section 2.7).

Assigning a probability to either a substantive or a statistical model is very different from asserting it is approximately correct or adequate for solving a problem. The philosopher of science Peter Achinstein had hoped to discover that his scientific heroes, Isaac Newton and John Stuart Mill, were Bayesian probabilists, but he was disappointed; what he finds is enlightening:

> Neither in their abstract formulations of inductive generalizations (Newton’s rule 3; Mill’s definition of ‘induction’) nor in their examples of particular inductions to general conclusions of the form ‘all As are Bs’ does the term ‘probability’ occur. Both write that from certain specific facts we can conclude general ones – not that we can conclude general propositions with probability, or that general propositions have a probability . . . From the inductive premises we simply conclude that the generalization is true, or as Newton allows in rule 4, ‘very nearly true,’ by which he appears to mean not ‘probably

true' but 'approximately true' (as he does when he takes the orbits of the satellites of Jupiter to be circles rather than ellipses). (Achinstein 2010, p. 176)

There are two main ways the "all models are false" charge comes about:

1. The statistical inference refers to an idealized and partial representation of a theory or process.
2. The probability model, to which a statistical inference refers, is at most an idealized and partial representation of the actual data-generating source.

Neither of these facts precludes the use of these *false* models to find out true things, or to correctly solve problems. On the contrary, it would be impossible to learn about the world if we did not deliberately falsify and simplify.

Adequacy for a Problem. The statistician George Box, to whom the slogan "all models are wrong" is often attributed, goes on to add "But some are useful" (1979, p. 202). I'll go further still: all models are false, no useful models are true. Were a model so complex as to represent every detail of data "realistically," it wouldn't be useful for finding things out. Let's say a statistical model is useful by being adequate for a problem, meaning it may be used to find true or approximately true solutions. Statistical hypotheses may be seen as conjectured solutions to a problem. A statistical model is adequate for a problem of statistical inference (which is only a subset of uses of statistical models) if it enables controlling and assessing if purported solutions are well or poorly probed, and to what degree. Through approximate models, we learn about the "important stable aspects" or systematic patterns when we are in the context of phenomena that exhibit statistical variability. When I speak of ruling out mistaken interpretations of data, I include mistakes about theoretical and causal claims. If you're an anti-realist about science, you will interpret, or rather reinterpret, theoretical claims in terms of observable claims of some sort. One such anti-realist view we've seen is instrumentalism: unobservables including genes, particles, light bending may be regarded as at most instruments for finding out about observable regularities and predictions. Fortunately we won't have to engage the thorny problem of realism in science, we can remain agnostic. Neither my arguments, nor the error statistical philosophy in general, turn on whether one adopts one of the philosophies of realism or anti-realism. Today's versions of realism and anti-realism are quite frankly too hard to tell apart to be of importance to our goals. The most important thing is that both realists and non-realists require an account of statistical inference. Moreover, whatever one's view of scientific theories, a statistical analysis of problems of actual experiments involves abstraction and creative analogy.

Testing Assumptions is Crucial. You might hear it charged that frequentist methods presuppose the assumptions of their statistical models, which is puzzling because when it comes to testing assumptions it's to frequentist methods that researchers turn.

> It is crucial that any account of statistical inference provides a conceptual framework for this process of model criticism, . . . the ability of the frequentist paradigm to offer a battery of simple significance tests for model checking and possible improvement is an important part of its ability to supply objective tools for learning. (Cox and Mayo 2010, p. 285)

Brad Efron is right to say the frequentist is the pessimist, who worries that "if anything can go wrong it will," while the Bayesian optimistically assumes if anything can go right it will (Efron 1998, p. 99). The frequentist error statistician is a worrywart, resigned to hoping things are half as good as intended. This also makes her an activist, deliberately reining in some portion of a problem so that it's sufficiently like one she knows how to check. Within these designated model checks, assumptions under test are intended to arise only as i-assumptions. They're assumptions for drawing out consequences, for possible falsification.

"In principle, the information in the data is split into two parts, one to assess the unknown parameters of interest and the other for model criticism" (Cox 2006a, p. 198). The number of successes in n Bernoulli trials, recall, is a *sufficient* statistic, and has a Binomial sampling distribution determined by θ, the probability of success on each trial (Section 3.3). If the model is appropriate then any permutation of the r successes in n trials has a known probability. Because this conditional distribution (X given s) is known, it serves to assess if the model is violated. If it shows statistical discordance, the model is disconfirmed or falsified. The key is to look at residuals: the difference between each observed value and what is expected under the model. (We illustrate with the runs test in Section 4.11.) It is also characteristic of error statistical methods to be relatively robust to violation.

Central Limit Theorem. In the presentation on justifying induction (Section 2.7), we heard Neyman stress how the empirical Law of Large Numbers (LLN) is in sync with the mathematical law in a number of "real random experiments." Supplementing the LLN is the Central Limit Theorem (CLT). It tells us that the mean $\overline{X}$ of n independent random variables, each X with mean μ, and finite non-zero σ^2, is approximately Normally distributed with its mean equal to μ and standard deviation $\sigma/\sqrt{n}$ – regardless of the underlying distribution of X. So long as n is reasonably large (say 40 or 50), and the underlying distribution is not too asymmetrical, the Normal distribution gives a good approximation, and is robust for many cases where IID is violated. The CLT tells us that $\overline{X}$ standardized is N(0,1). The finite non-zero variance isn't much of a restriction, and even this has been capable of being relaxed.

The CLT links a claim or question about a statistical hypothesis to claims about the relative frequencies that would be expected in applications (real or hypothetical) of the experiment. Owing to this link, we can use the sample mean to inquire about values of μ that are capable or incapable of bringing it about. Our standardized difference measures observed disagreement, and classifies those improbably far from hypothesized values. Thus, statistical models may be adequate for real random experiments, and hypotheses to this effect may pass with severity.

Exhibit (xii): Pest Control. Neyman (1952) immediately turns from the canonical examples of real random experiments – of coin tossing and roulette wheels – to illustrate how "the abstract theory of probability . . . may be, and actually is, applied to solve problems of practice importance" such as pest control (p. 27)! Given the lack of human control here, he expects the mechanism to be complicated. The first attempt to model the variation in larvae hatched from moth eggs is way off.

> [I]f we attempt to treat the distribution of larvae from the point of view of [the Poisson distribution], we would have to assume that each larva is placed on the field independently of the others. This basic assumption was flatly contradicted by the life of larvae as described by Dr. Beall. Larvae develop from eggs laid by moths. It is plausible to assume that, when a moth feels like laying eggs, it does not make any special choice between sections of a field planted with the same crop and reasonably uniform in other respects. (1952, pp. 34–5).

So it's plausible to suppose the Poisson distribution for the spots where moths lay their eggs. However, a data analysis made it "clear that a very serious divergence exists" between the actual distribution of larvae and the Poisson model (ibid., p. 34). Larvae expert, Dr. Beall, explains why: At each "sitting" a moth lays a whole batch of eggs and the number of eggs varies from one cluster to another. "After hatching . . . the larvae begin to look for food and crawl around" but given their slow movement "if one larva is found, then it is likely that the plot will contain more than one from the same cluster" (ibid., p. 35). An independence assumption fails. (I omit many details; see Neyman 1952, Gillies 2001.)

The main thing is this: The misfit with the Poisson model leads Neyman to arrive at a completely novel distribution: he called it the type A distribution (a "contagious" distribution). Yet Neyman knows full well that even the type A distribution is strictly inadequate, and a far more complex distribution would be required for answering certain questions. He knows it's strictly false. Yet it suffices to show why the first attempt failed, and it's adequate to solving his immediate problem in pest control.

Souvenir U: Severity in Terms of Problem-Solving

The aim of inquiry is finding things out. To find things out we need to solve problems that arise due to limited, partial, noisy, and error-prone information. Statistical models are at best approximations of aspects of the data-generating process. Reasserting this fact is not informative about the case at hand. These models work because they need only capture rather coarse properties of the phenomena: the error probabilities of the test method are approximately and conservatively related to actual ones. A problem beset by variability is turned into one where the variability is known at least approximately. Far from wanting true (or even "truer") models, we need models whose deliberate falsity enables finding things out.

Our threadbare array of models and questions is just a starter home to grow the nooks and crannies between data and what you want to know (Souvenir E, Figure 2.1). In learning about the large-scale theories of sexy science, intermediate statistical models house two "would-be" claims. Let me explain. The theory of GTR does not directly say anything about an experiment we could perform. Splitting off some partial question, say about the deflection effect, we get a prediction about what *would be* expected were the deflection effect approximately equal to the Einstein value, 1.75". Raw data from actual experiments, cleaned and massaged, afford inferences about intermediate (astrometric) models; inferences as to what it would be like were we taking measurements at the limb of the sun. The two counterfactual inferences – from the theory down, and the data up – meet in the intermediate statistical models. We don't seek a probabilist assignment to a hypothesis or model. We want to know what the data say about a conjectured solution to a problem: What erroneous interpretations have been well ruled out? Which have not even been probed? The warrant for these claims is afforded by the method's capabilities to have informed us of mistaken interpretations. *Statistical methods are useful for testing solutions to problems when this capability/incapability is captured by the relative frequency with which the method avoids misinterpretations.*

If you want to avoid speaking of "truth" you can put the severity requirement in terms of solving a problem. A claim *H* asserts a proposed solution S to an inferential problem is adequate in some respects. It could be a model for prediction, or anything besides.

H: S is adequate for a problem

To reject *H* means "infer *S* is inadequate for a problem." If none of the possible outcomes lead to reject *H* even if *H* is false – the test is incapable of finding inadequacies in *S* – then "do not reject *H*" is BENT evidence that *H* is true. We move from no capability, to some, to high:

> If the test procedure (which generally alludes to a cluster of tests) very rarely rejects *H*, if *H* is true, then "reject *H*" provides evidence for falsifying *H* in the respect indicated.

You could say, a particular inadequacy is corroborated. It's still an inferential question: what's warranted to infer. We start, not with hypotheses, but questions and problems. We want to appraise hypothesized answers severely.

I'll meet you in the ship's library for a reenactment of George Box (1983) issuing "An Apology for Ecumenism in Statistics."

4.9 For Model-Checking, They Come Back to Significance Tests

> Why can't all criticism be done using Bayes posterior analysis . . .? The difficulty with this approach is that by supposing all possible sets of assumptions known *a priori*, it discredits the possibility of new discovery. But new discovery is, after all, the most important object of the scientific process. (George Box 1983, p. 73)

Why the apology for ecumenism? Unlike most Bayesians, Box does not view induction as probabilism in the form of probabilistic updating (posterior probabilism), or any form of probabilism. Rather, it requires critically testing whether a model M_i is "consonant" with data, and this, he argues, demands frequentist significance testing. Our ability "to find patterns in discrepancies $M_i - y_\mathrm{d}$ between the data and what might be expected if some tentative model were true is of great importance in the search for explanations of data and of discrepant events" (Box 1983, p. 57). But the dangers of apophenia raise their head.

> However, some check is needed on [the brain's] pattern seeking ability, for common experience shows that some pattern or other can be seen in almost any set of data or facts. This is the object of diagnostic checks and tests of fit which, I will argue, require frequentist theory significance tests for their formal justification. (ibid.)

Once you have inductively arrived at an appropriate model, the move, on his view, "is entirely *deductive* and will be called *estimation*" (ibid., p. 56). The deductive portion, he thinks, can be Bayesian, but the inductive portion requires frequentist significance tests, and statistical inference depends on an iteration between the two. Alluding to Box, Peter Huber avers: "Within orthodox Bayesian statistics, we cannot even address the question whether a model M_i, under consideration at stage i of the investigation, is consonant with the data y" (Huber 2011, p. 92). Box adds a non-Bayesian activity to his account.

A note on Box's slightly idiosyncratic use of deduction/induction: Frequentist significance testing is often called deductive, but for Box it's the inductive component. There's no confusion if we remember that Box is emphasizing that frequentist testing is the source of new ideas, it is the inductive achievement. It's in sync with our own view that inductive inference to claim *C* consists of trying and failing to falsify *C* with a stringent test: *C* should be well corroborated. In fact, the approach to misspecification (M-S) testing that melds seamlessly with the error statistical account has its roots in the diagnostic checking of Box and Jenkins (1976).

All You Need Is Bayes. Not

Box and Jenkins highlight the link between 'prove' and 'test': "A model is only capable of being 'proved' in the biblical sense of being put to the test" (ibid., p. 286). Box considers the possibility that model checking occurs as follows: One might imagine A_1, A_2, …, A_k being alternative assumptions and then computing $\Pr(A_i|y)$. Box denies this is plausible. To assume we start out with all models precludes the "something else we haven't thought of" so vital to science. Typically, Bayesians try to deal with this by computing a Bayesian catchall "everything else." Savage recommends reserving a low prior for the catchall (1962a), but Box worries that this may allow you to assign model M_i a high posterior probability *relative* to the other models considered. "In practice this would seem of little comfort" (Box 1983, pp. 73–4). For suppose of the three models under consideration the posteriors are 0.001, 0.001, 0.998, but unknown to the investigator a fourth model is a thousand times more probable than even the most probable one considered so far.

So he turns to frequentist tests for model checking. Is there any harm in snatching some cookies from the frequentist cookie jar? Not really. Does it violate the Likelihood Principle (LP)? Let's listen to Box:

> The likelihood principle holds, of course, for the estimation aspect of inference in which the model is temporarily assumed true. However it is inapplicable to the criticism process in which the model is regarded as in doubt … In the criticism phase we are considering whether, given A, the sample y_d is likely to have occurred at all. To do this

> we *must* consider it in relation to the *other* samples that could have occurred but did not. (Box 1983, pp. 74–5)

Suppose you're about to use a statistical model, say n IID Normal trials for a primary inference about mean μ. Checking independence (I), identical distributed (ID), or the Normality assumption (N) are *secondary inferences* in relation to the primary one. In conducting secondary inferences, Box is saying, the LP must be violated, or simply doesn't apply. You can run a simple Fisherian significance test – the null asserting the model assumption A holds – and reject it if the observed result is statistically significantly far from what A predicts. A *P*-value (or its informal equivalent) is computed – a tail area – which requires considering outcomes other than the one observed.

Box gives the example of stopping rules. Stopping rules don't alter the posterior distribution, as we learned from the extreme example in Excursion 1 (Section 1.5). For a simple example, he considers four Bernoulli trials: ⟨S, S, F, S⟩. The same string could have come about if $n = 4$ was fixed in advance, or if the plan was to sample until the third success is observed. The latter are called negative Binomial trials, the former Binomial. The string enters the likelihood ratio the same way, $\begin{pmatrix} 4 \\ 3 \end{pmatrix}\theta^3(1-\theta)$ and $\begin{pmatrix} 3 \\ 2 \end{pmatrix}\theta^3(1-\theta)$ respectively: the only difference is the coefficients, which cancel. But the significance tester distinguishes them, because the sample space, and corresponding error probabilities, differ.[1] When it comes to model testing, Box contends, this LP violation is altogether reasonable, since "we are considering whether, given A, the sample is likely to have occurred at all" (ibid., p. 75).

This is interesting. Isn't it also our question at his estimation stage where the LP denies stopping rules matter? We don't know there's any genuine effect, or if a null is true. If we ignore the stopping rules, we may make it too easy to find one, even if it's absent. In the example of Section 1.5, we ensure erroneous rejection, violating "weak repeated sampling." A Boxian Bayesian, who retains the LP for primary statistical inference, still seems to owe us an explanation why we shouldn't echo Armitage (1962, p. 72) that "Thou shalt be misled" if your method hides optional stopping at the primary (Box's estimation) stage.

Another little puzzle arises in telling what's true about the LP: Is the LP violated or simply inapplicable in secondary testing of model assumptions. Consider Casella and R. Berger's text.

> Most data analysts perform some sort of 'model checking' when analyzing a set of data . . . For example, it is common practice to examine *residuals* from a model, statistics that

[1] The sufficient statistic in the negative Binomial case is *N*, the number of trials until the fourth success. In the Binomial case, it is $\overline{X}$ (Cox and Mayo 2010, p. 286).

> measure variation in the data not accounted for by the model . . . (Of course such a practice directly violates the Likelihood Principle also.) Thus, *before* considering [the Likelihood Principle], we must be comfortable with the model. (Casella and R. Berger 2002, pp. 295–6)

For them, it appears, the LP is full out violated in model checking. I'm not sure how much turns on whether the LP is regarded as violated or merely inapplicable in testing assumptions; a question arises in either case. Say you have carried out Box's iterative moves between criticism and estimation, arrived at a model deemed adequate, and infer H: model M_i is adequate for modeling data $\boldsymbol{x}_0$. My question is: How is this secondary inference qualified? Probabilists are supposed to qualify uncertain claims with probability (e.g., with posterior probabilities or comparisons of posteriors). What about this secondary inference to the adequacy/inadequacy of the model? For Boxians, it's admitted to be a non-Bayesian frequentist animal. Still a long-run performance justification wouldn't seem plausible. If you're going to accept the model as sufficiently adequate to build the primary inference, you'd want to say it had passed a severe test: that if it wasn't adequate for the primary inference, then you probably would have discovered this through the secondary model checking. However, if secondary inference is also a statistical inference, it looks like Casella and R. Berger, and Box, are right to consider the LP violated – as regards *that inference*. There's an appeal to outcomes other than the one observed.

Andrew Gelman's Bayesian approach can be considered an offshoot of Box's, but, unlike Box, he will avoid assigning a posterior probability to the primary inference. Indeed, he calls himself a falsificationist Bayesian, and is disgruntled that Bayesians don't test their models.

> I vividly remember going from poster to poster at the 1991 Valencia meeting on Bayesian statistics . . . not only were they not interested in checking the fit of the models, they considered such checks to be illegitimate. To them, any Bayesian model necessarily represented a subjective prior distribution and as such could never be tested. The idea of testing and p-values were held to be counter to the Bayesian philosophy. (Gelman 2011, pp. 68–9)

What he's describing is in sync with the classical subjective Bayesian: If "the Bayesian theory is about coherence, not about right or wrong" (Lindley 1976, p. 359), then what's to test? Lindley does distinguish a pre-data model choice:

> Notice that the likelihood principle only applies to inference, i.e. to calculations once the data have been observed. Before then, e.g. in some aspects of model choice, in the design of experiments . . ., a consideration of several possible data values is essential. (Lindley 2000, p. 310)

This he views as a decision based on maximizing an agent's expected utility. But wouldn't a correct assessment of utility depend on information on model adequacy?

Interestingly, there are a number of Bayesians who entertain the idea of a Bayesian *P*-value to check accordance of a model when there's no alternative in sight.[2] They accept the idea that significance tests and *P*-values are a good way, if not the only way, to assess the consonance between data and model. Yet perhaps they are only grinning and bearing it. As soon as alternative models are available, most would sooner engage in a Bayesian analysis, e.g., Bayes factors (Bayarri and Berger 2004).

But Gelman is a denizen of a tribe of Bayesians that rejects these traditional forms. "To me, Bayes factors correspond to a discrete view of the world, in which we must choose between models A, B, or C" (Gelman 2011, p. 74) or a weighted average of them as in Madigan and Raftery (1994). Nor will it be a posterior. "I do not trust Bayesian induction over the space of models because the posterior probability of a continuous-parameter model depends crucially on untestable aspects of its prior distribution" (Gelman 2011, p. 70). Instead, for Gelman, the priors/posteriors arise as an interim predictive device to draw out and test implications of a model. What is the status of the inference to the adequacy of the model? If neither probabilified nor Bayes ratioed, it can at least be well or poorly tested. In fact, he says, "This view corresponds closely to the error-statistics idea of Mayo (1996)" (ibid., p. 70). We'll try to extricate his approach in Excursion 6.

4.10 Bootstrap Resampling: My Sample Is a Mirror of the Universe

"My difficulty" with the Likelihood Principle (LP), declares Brad Efron (in a comment on Lindley), is that it "rules out many of our most useful data analytic tools without providing workable substitutes" (2000, p. 330) – notably, the method for which he is well known: bootstrap resampling (Efron 1979). Let's take a little detour to have a look around this hot topic. (I follow D. Freedman (2009), and A. Spanos (201~~8~~)).

We have a single IID sample of size 100 of the water temperatures soon after the accident $\boldsymbol{x}_0 = \langle x_1, x_2, \ldots, x_{100} \rangle$. Can we say anything about its accuracy even if we couldn't take any more? Yes. We can lift ourselves up by the bootstraps with this single $\boldsymbol{x}_0$ by treating it as its own population. Get the computer to take

[2] There is considerable discussion as to which involve pejorative "double use" of data, and which give adequate frequentist guarantees or calibrations (Ghosh et al. 2006, pp. 175–84; Bayarri and Berger 2004). But I find their rationale unclear.

a large number, say 10,000, independent samples from $\boldsymbol{x}_0$ (with replacement), giving 10,000 *resamples*. Then reestimate the mean for each, giving 10,000 bootstrap means $\overline{X}_{\mathrm{b}_1}, \overline{X}_{\mathrm{b}_2}, \ldots, \overline{X}_{\mathrm{b}_{10,000}}$. The frequency with which the bootstrapped means take different values approximates the sampling distribution of $\overline{X}$. It can be extended to medians, standard deviations, etc. "This is exactly the kind of calculation that is ruled out by the likelihood principle; it relies on hypothetical data sets different from the data that are actually observed and does so in a particularly flagrant way" (Efron 2000, p. 331). At its very core is the question: what would mean temperatures be like were we to have repeated the process many times? This lets us learn: How capable of producing our observed sample is a universe with mean temperature no higher than the temperature thought to endanger the ecosystem?

Averaging the 10,000 bootstrap means, we get the overall bootstrap sample mean, $\overline{X}_{\mathrm{b}}$. If n is sufficiently large, the resampling distribution of $\overline{X}_{\mathrm{b}} - \overline{x}$ mirrors the sampling distribution of $\overline{X} - \mu$, where μ is the mean of the population. We can use the sample deviation of $\overline{X}_{\mathrm{b}}$ to approximate the standard error of $\overline{X}$.[3]

To illustrate with a tiny example, imagine that instead of 100 temperature measurements there are only 10: $\boldsymbol{x}_0$: 150, 165, 151, 138, 148, 167, 164, 160, 136, 173, with sample mean $\overline{x} = 155.2$, and instead of 10,000 resamples, only 5. Since it's with replacement there can be duplicates.

$\boldsymbol{x}_0$: 150, 165, 151, 138, 148, 167, 164, 160, 136, 173	$\overline{x} = 155.2$
Bootstrap resamples	**Bootstrap means**
$\boldsymbol{x}_{\mathrm{b}_1}$: 160, 136, 138, 165, 173, 165, 167, 148, 151, 167	157
$\boldsymbol{x}_{\mathrm{b}_2}$: 164, 136, 165, 167, 148, 138, 151, 160, 150, 151	153
$\boldsymbol{x}_{\mathrm{b}_3}$: 173, 138, 173, 160, 167, 167, 148, 138, 148, 165	157.7
$\boldsymbol{x}_{\mathrm{b}_4}$: 148, 138, 164, 167, 160, 150, 164, 167, 148, 173	157.9
$\boldsymbol{x}_{\mathrm{b}_5}$: 173, 136, 167, 138, 150, 160, 148, 164, 164, 148	154.8

Here are the rest of the bootstrap statistics:

Bootstrap overall mean: $\overline{x}^{\mathrm{b}} = [157 + 153 + 157.7 + 157.9 + 154.8]/5 = 156.08$;
Bootstrap variance: $[(157 - 156.08)^2 + (153 - 156.08)^2 + (157.7 - 156.08)^2 + (157.9 - 156.08)^2 + (154.8 - 156.08)^2]/4 = 4.477$;
Bootstrap SE: $\sqrt{}4.477 = 2.116$.

Note the difference between the mean of our observed sample and that of the overall bootstrap mean (the bias) is small: $\left(\overline{x} - \overline{x}^{\mathrm{b}}\right) = 155.2 - 156.08 = -0.88$.

[3] The bootstrapped distribution is conditional on the observed $\boldsymbol{x}$.

From our toy example, we could form the bootstrap 0.95 confidence interval: 156.08 ± 1.96 (2.116) approximately [152, 158]. You must now imagine we arrived at the interval via 10,000 samples, not 5. The observed mean just after the accident (155.2) exceeds 150 by around 2.5 SE, indicating our sample came from a population with $\theta > 150$. In fact, were $\theta \leq 152$, such large results would occur infrequently.

The non-parametric bootstrap works without relying on a theoretical probability distribution, at least when the sample is random, large enough, and has sufficiently many bootstraps. Statistical inference by non-parametrics still has assumptions, such as IID (although there are other variants). Many propose we do all statistics non-parametrically, and some newer texts advocate this. I'm all for it, because the underlying error statistical reasoning becomes especially clear. I concur, too, with the philosopher Jan Sprenger that philosophy of statistics should integrate "resampling methods into a unified scheme of data analysis and inductive inference" (Sprenger 2011, p. 74). That unified scheme is error statistical. (I'm not sure how it's in sync with his subjective Bayesianism.)

The philosophical significance of bootstrap resampling is twofold. (1) The relative frequency of different values of $\overline{X}_b$ sustains our error statistical argument: the probability model is a good way to approximate the empirical distribution analytically. Through a hypothetical – 'what it would be like' were this process repeated many times – we understand what produced the single observed sample. (2) By identifying exemplary cases where we manage to take approximately random samples, we can achieve inductive lift-off. It's through our deliberate data generation efforts, in other words, that we solve induction. I don't know if taking water samples is one such exemplar, but I'm using it just as an illustration. We may imagine ample checks of water sampling on bodies with known temperature show we're pretty good at taking random samples of water temperature. Thus we reason, it works when the mean temperature is unknown. Can supernal powers read my mind and interfere just in the cases of an unknown mean?

> Nor is it necessary to deny altogether the existence of mysterious influences adverse to the validity of the inductive . . . processes. So long as their influence were not too overwhelming, the wonderful self-correcting nature of the ampliative inference would enable us . . . to detect and make allowance for them. (Peirce 2.749)

4.11 Misspecification (M-S) Testing in the Error Statistical Account

Induction – understood as severe testing – "not only corrects its conclusions, it even corrects its premises" (Peirce 3.575). In the land of statistical inference it does so by checking and correcting the assumptions underlying the inference. It's common to distinguish "model-based" and "design-based" statistical inference, but both involve assumptions. So let's speak of the adequacy of the model in both cases. It's to this auditing task that I now turn. Let's call violated statistical assumptions statistical model misspecifications. The term "misspecification" is often used to refer to a problem with a primary model, whereas for us it will always refer to the secondary problem of checking assumptions for probing a primary question (following A. Spanos). "Primary" is relative to the main inferential task: Once an adequate statistical model is at hand, an inquiry can grow to include many layers of primary questions.

Splitting things off piecemeal has payoffs. The key is for the relevant error probabilities to be sufficiently close to those calculated in probing the primary claim. Even if you have a theory, turning it into something statistically testable isn't straightforward. You can't simply add an error term at the end, such as y = theory + error, particularly in the social sciences – although people often do. The trouble is that you can tinker with the error term to "fix" anomalies without the theory having been tested in the least. Aris Spanos, an econometrician, roundly criticizes this tendency to "the preeminence of theory in econometric modeling" (2010c, p. 202). This would be okay if you were only estimating quantities in a theory known to be true or adequate, but in fact, Spanos says, "mainstream economic theories have been invariably unreliable predictors of economic phenomena" (ibid., p. 203).

As always, different inquiry types have their own error repertoires that need to be mastered. Allow me to try to climb neatly through the vegetation of one of the more treacherous modeling fields: econometrics. Relying on seven figures from Mayo and Spanos (2004), I'll tell the story of a case Spanos presented to me in 2002.

Nonsense Regression

Suppose that in her attempt to find a way to understand and predict changes in the US population, an economist discovers an empirical relationship that appears to provide almost a "law-like" fit:

$$y_t = 167 + 2x_t + \hat{u}_t,$$

where y_t denotes the population of the USA (in millions), and x_t denotes a secret variable whose identity Spanos would not reveal until the end of the analysis. The subscript t is time. There are 33 annual data points for the period 1955–1989 ($t = 1$ is 1955, $t = 2$ is 1956, etc.) The data can be represented as 33 pairs $\mathbf{z}_0 = \{(x_t, y_t), t = 1, 2, \ldots, 33\}$. The coefficients 167 and 2 come from the least squares fit, a purely mathematical operation.

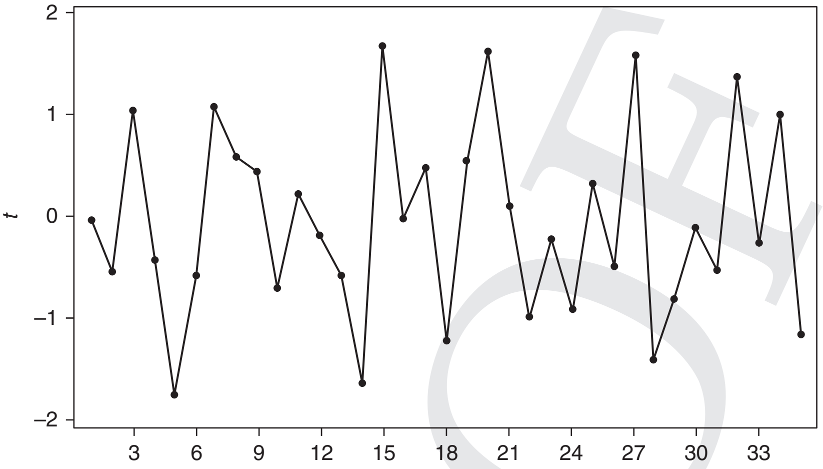

Figure 4.2 A typical realization of a NIID process.

This is an example of fitting a *Linear Regression Model* (LRM), which forms the backbone of most statistical models of interest:

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

The term $\beta_0 + \beta_1 x_t$ is viewed as the *systematic* component (and is the expected value of y_t), and $u_t = y_t - \beta_0 - \beta_1 x_t$ is the error or *non-systematic* component. The error u_t is a random variable assumed to be Normal, Independent, and Identically Distributed (NIID) with mean 0, variance σ^2. This is called Normal white noise. Figure 4.2 shows what NIID looks like.

A Primary Statistical Question: How Good a Predictor Is x_t? The empirical equation is intended to enable us to understand how y_t varies with x_t. Testing the statistical significance of the coefficients shows them to be highly significant: *P*-values are zero (0) to a third decimal, indicating a very strong relationship between the variables. The goodness-of-fit measure of how well this model “explains” the variability of y_t, $R^2 = 0.995$, an almost perfect fit (Figure 4.3). Everything looks hunky dory. Is it reliable? Only if the errors are approximately NIID with mean 0, variance σ^2.

The null hypotheses in M-S tests take the form

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

as against not-H_0, which, strictly speaking, would include all of the ways one or more of its assumptions can fail. To rein in the testing, we consider specific departures with appropriate choices of test statistic d($\boldsymbol{y}$).

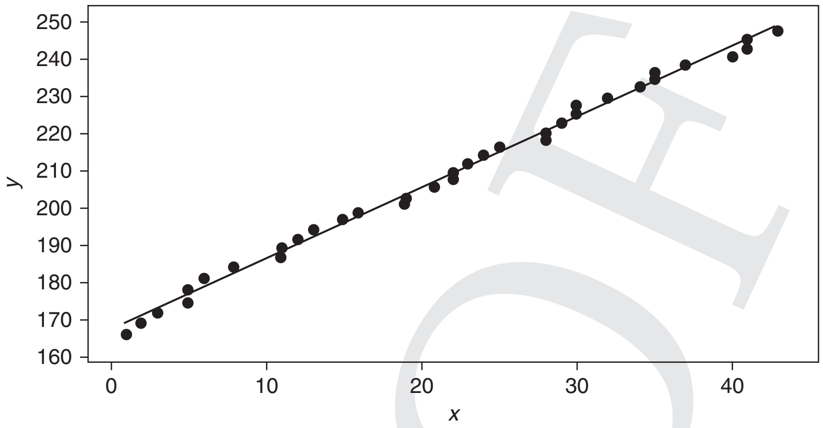

Figure 4.3 Fitted line plot, $y = 167.1 + 1.907x$.

Residuals Are the Key

Testing Randomness. The non-parametric *runs test* for IID is the test I showed Wes Salmon in relation to the Bernoulli model and justifying induction (Section 2.7). It falls under "omnibus" tests in Cox's taxonomy (Section 3.3). It can apply here by re-asking our question regarding the LRM. Look at the graph of the residuals (Figure 4.4), where the "hats" are the fitted values for the coefficients:

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

If the residuals do not fluctuate like pure noise, it's a sign the sample is not IID. Instead of the value of each residual, record whether the difference between successive observations is positive (+) or negative (–),

$$\underbrace{++}_{1}\ \underbrace{-}_{2}\ \underbrace{++}_{3}\ \underbrace{-}_{4}\ \underbrace{+++}_{5}\ \underbrace{-}_{6}\ \underbrace{+}_{7}\ \underbrace{-}_{8}\ \underbrace{+}_{9}\ \underbrace{--}_{10}\ \underbrace{+++++}_{11}\ \underbrace{-}_{12}\ \ldots$$

Each sequence of pluses only, or minuses only, is a *run*. We can calculate the probability of the number of runs just from the hypothesis that the assumption of randomness holds. It serves only as an argumentative (or i) assumption for the check. The expected number of runs, under randomness, is $(2n - 1)/3$, or in our case of $n = 35$ values, 23. Running out of letters, I'll use R again for the number of runs. The distribution of the test statistic, $\tau(\boldsymbol{y}) = [R - E(R)]/\sqrt{\mathrm{Var}(R)}$, under IID for $n \geq 20$, can be approximated by

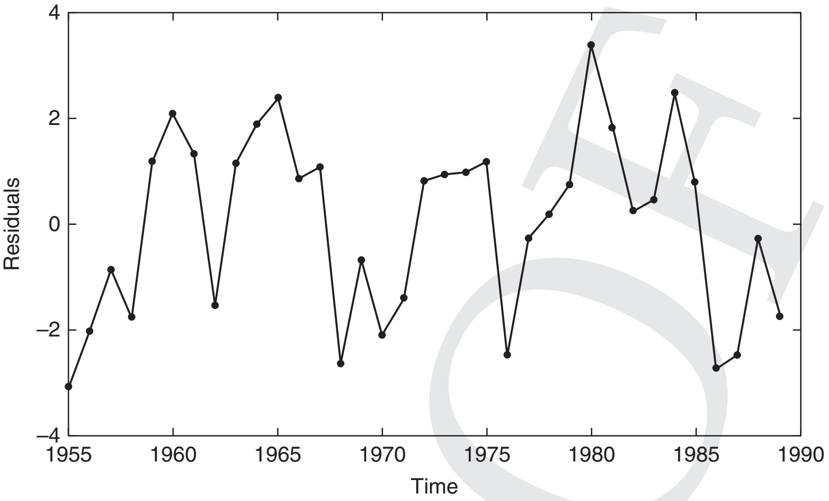

Figure 4.4 Plot of residuals over time.

N(0, 1). We're actually testing H_0: $E(R) = (2n - 1)/3$ vs. H_1: $E(R) \neq (2n- 1)/3$, E is the expected value. We reject H_0 iff the observed R differs sufficiently (in either direction) from $E(R) = 23$. (SE = $\sqrt{}(16n–29)/90$.)

Our data yield 18 runs, around 2.4 SE units, giving a P-value of approximately 0.02. So 98% of the time, we'd expect an R closer to 23 if IID holds. Arguing from severity, the data indicate non-randomness. But rejecting the null only indicates a denial of IID: either independence is a problem or identically distributed is a problem. The test itself does not indicate whether the fault lies with one or the other or both. More specific M-S testing enables addressing this Duhemian problem.

The Error in Fixing Error. A widely used parametric test for independence is the Durbin–Watson (D-W) test. Here, all the assumptions of the LRM are retained, except the one under test, independence, which is "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 own past, that is, *autocorrelated*,

$$u_t = \rho u_{t-1} + \varepsilon_t, \;\; t = 1, 2, \ldots, n, \ldots$$

This is to propose a new overarching model, the *Autocorrelation-Corrected LRM*:

$$\text{Proposed } M_1\text{:}\;\; y_t = \beta_0 + \beta_1 x_t + u_t, \quad u_t = \rho u_{t-1} + \varepsilon_t, \quad t = 1, 2, \ldots, n, \ldots$$

(Now it's ε_t that is assumed to be a Normal, white noise process.) The D-W test assesses whether or not $\rho = 0$, *assuming we are within model* M_1. One way to bring this out is to view the D-W test as actually considering the conjunctions:

$$H_0: \{M_1 \;\&\; \rho = 0\} \text{ vs. } H_1: \{M_1 \;\&\; \rho \neq 0\}.$$

With the data in our example, the D-W test statistic rejects the null hypothesis (at level 0.02), which is standardly taken as grounds to adopt H_1. This is a mistake. This move, to infer H_1, is warranted only if we are within M_1. True, 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. Notice we are in one of the "non-exhaustive" pigeonholes ("nested") of Cox's taxonomy. 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 $\rho \neq 0$. Thus any inference to H_1 lacks severity. The resulting model will *appear* to have corrected for autocorrelation even when it is statistically inadequate. If used for the "primary" statistical inferences, the actual error probabilities are much higher than the ones it is thought to license, and such inferences are unreliable at predicting values beyond the data used. "This is the kind of cure that kills the patient," Spanos warns. What should we do instead?

Probabilistic Reduction: Spanos

Spanos shows that any statistical model can be specified in terms of probabilistic assumptions from three broad categories: Distribution, Dependence, and Heterogeneity. In other words, a model emerges from selecting probabilistic assumptions from a menu of three groups: a choice of distribution; of type of dependence, if any; and a type of heterogeneity, i.e., how the generating mechanism remains the same or changes over the ordering of interest, such as time, space, or individuals. The LRM reflects just one of many ways of reducing the set of all possible models that could have given rise to the data $\boldsymbol{z}_0 = \{(x_t, y_t), t = 1, \ldots, n\}$: Normal, Independent, Identically Distributed (NIID). Statistical inference need not be hamstrung by the neat and tidy cases. As a first step, we partition the set of all possible models coarsely:

	Distribution	**Dependence**	**Heterogeneity**
LRM	Normal	Independent	Identically distributed
Alternative (coarse partition)	Non-normal	Dependent	Non-IID

Since we are partitioning or reducing the space of models by means of the probabilistic assumptions, Spanos calls it the *Probabilistic Reduction* (PR)

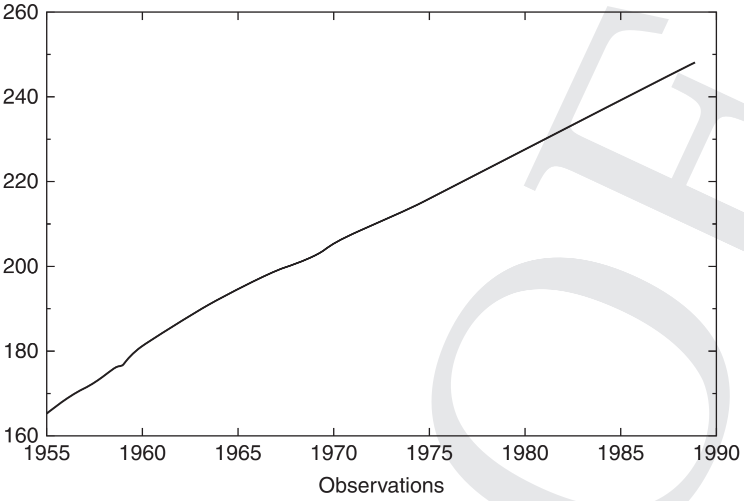

Figure 4.5 USA population (y) over time.

approach (first in Spanos 1986, 1999). The PR approach to M-S testing weaves together threads from Box–Jenkins, and what some dub the LSE (London School of Economics) tradition. Rather than give the assumptions by means of the error term, as is traditional, Spanos will specify them in terms of the observable random variables (x_t, y_t). This has several advantages. For one thing, it brings out hidden assumptions, notably assuming the parameters (β_0, β_1, σ^2) do not change with t. This is called *t-homogeneity* or t-invariance. Second, we can't directly probe errors given by the error term, but we can indirectly test them from the data.

We ask: What would be expected if each data series were to have come from a NIID process? Compare a typical realization of a NIID process (Figure 4.2) with the two series (t, x_t) and (t, y_t), in Figures 4.5 and 4.6, called t-plots.

Clearly, neither data series looks like the NIID of Figure 4.2. In each case the mean is increasing with time – there's a strong upward trend. Econometricians never use a short phrase when a long one will do, they call the trending mean: *mean heterogeneity*. The data can't be viewed as a realization of identically distributed random variables: ID is false. The very assumption of linear correlation between x and y is that x has a mean μ_x, and y has mean μ_y. If these are changing over the different samples, your estimate of correlation makes no sense.

We respecify, by adding terms of form t, t^2, . . ., to the model M_0. We don't know how far we'll have to go. We aren't inferring anything yet, just building a statistical model whose adequacy for the primary statistical inference will be

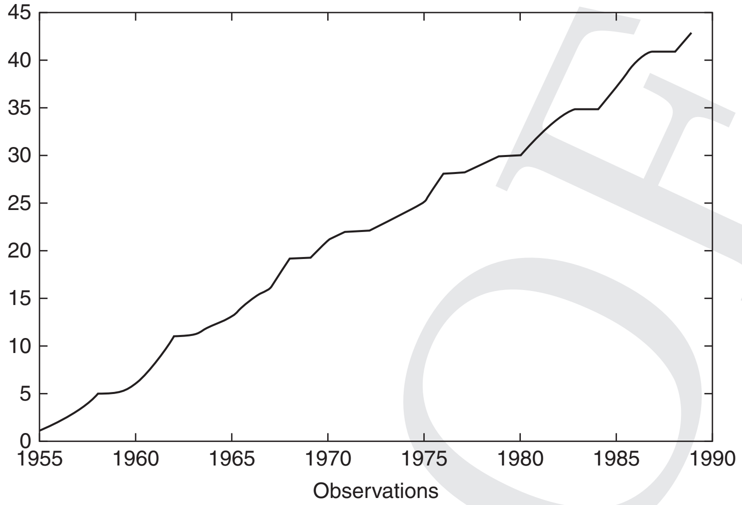

Figure 4.6 Secret variable (x) over time.

tested in its own right. With these data, adding t suffices to capture the trend, but in building the model you may include higher terms, allowing some to be found unnecessary later on. We can summarize our progress in detecting potential departures from the LRM model assumptions thus far:

	Distribution	Dependence	Heterogeneity
LRM	Normal	Independent	Identically distributed
Alternative	?	?	Mean heterogeneity

What about the independence assumption? We could check dependence if our data were ID and not obscured by the influence of the trending mean. We can "subtract out" the trending mean in a generic way to see what it would be like without it. Figures 4.7 and 4.8 show the *detrended* x_t and y_t. Reading data plots, and understanding how they connect to probabilistic assumptions, is a key feature of the PR approach.

The detrended data in both figures indicate, to a trained eye, positive dependence or "memory" in the form of cycles – this is called Markov dependence. So the independence assumption also looks problematic, and this explains the autocorrelation detected by the Durbin–Watson test and runs tests. As with trends, it comes in different orders, depending on how long the memory is found to be. It is modeled by adding terms called lags. To y_t add y_{t-1}, y_{t-2}, . . ., as many as needed. Likewise to x_t add x_{t-1}, x_{t-2}, . . . Our assessment so far, just on the basis of the graphical analysis is:

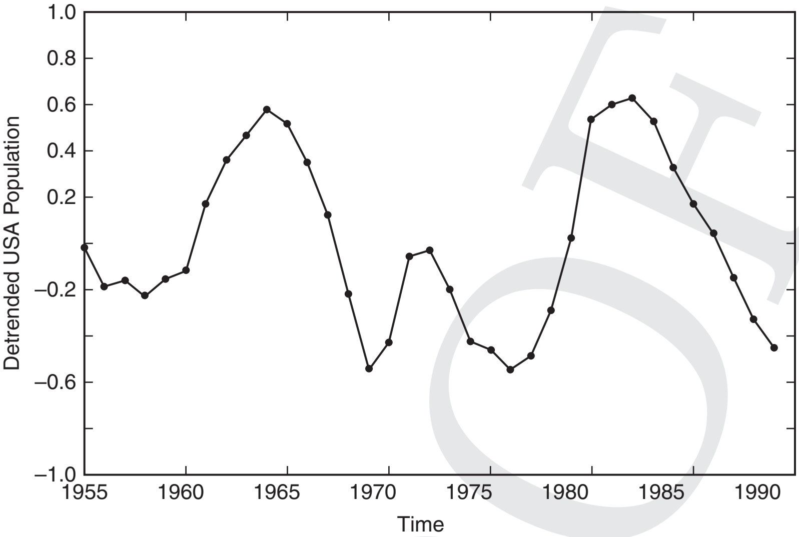

Figure 4.7 Detrended population data.

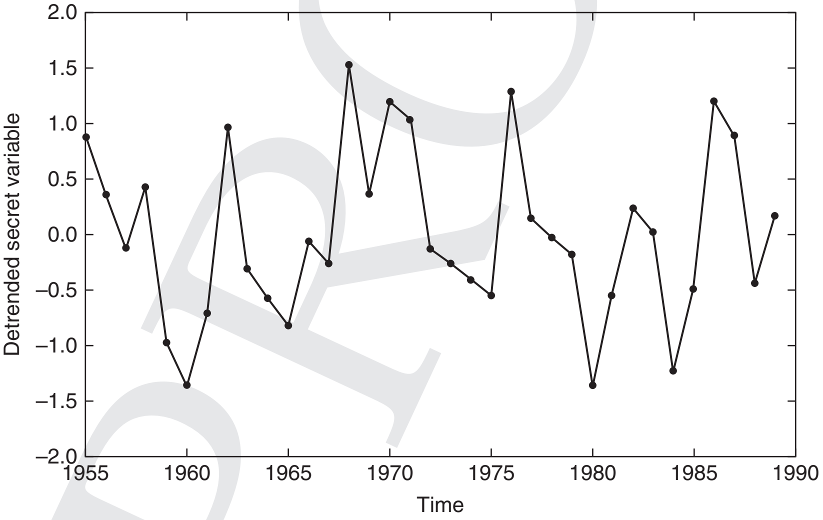

Figure 4.8 Detrended secret variable data.

	Distribution	Dependence	Heterogeneity
LRM	Normal	Independent	Identically distributed
Alternative	?	Markov	Mean heterogeneity

Finally, if we can see what the data $\boldsymbol{z}_0 = \{(x_t, y_t), t = 1, 2, \ldots, 35\}$ would look like without the heterogeneity ("detrended") and without the dependence ("dememorized"), we could get some ideas about the appropriateness of the Normality assumption. We do this by subtracting them out "on paper" again (shown on graphs). The scatter-plot of (x_t, y_t) shows the elliptical pattern expected for Normality (though I haven't included a figure). We can organize our respecified model as an alternative to the LRM.

	Distribution	Dependence	Heterogeneity
LRM	Normal	Independent	Identically distributed
Alternative	Normal	Markov	Mean heterogeneity

While there are still several selections under each of the menu headings of Markov dependence and mean heterogeneity, the length of the Markov dependence (m), and the degree (ℓ) of the polynomial in t, I'm imagining we've carried out the subsequent rounds of the probing strategy.

The model derived by re-partitioning the set of all possible models, using the new reduction assumptions of Normality, Markov, and mean heterogeneity is called the Dynamic Linear Regression Model (DLRM). These data require one trend and two lags:

$$\begin{aligned} M_2\colon\ y_t = \beta_0 + \beta_1 x_t + ct \text{ [trending mean]} + (\gamma_1 y_{t-1} + \gamma_2 y_{t-2}) \\ + (\gamma_3 x_{t-1} + \gamma_4 x_{t-2}) \text{ [temporal lags]} + \varepsilon_t. \end{aligned}$$

Back to the Primary Statistical Inference. Having established the statistical adequacy of the respecified model M_2 we are then licensed in making primary statistical inferences about the values of its parameters. In particular, does the secret variable help to predict the population of the USA (y_t)? No. A test of joint significance of the coefficients of (x_t, x_{t-1}, x_{t-2}), yields a P-value of 0.823 (using an F test). We cannot reject the hypothesis that they are all 0, indicating that x contributed nothing towards predicting or explaining y. The regression between x_t and y_t suggested by models M_0 and M_1 turns out to be a spurious or nonsense regression. Dropping the x variable from the model and reestimating the parameters we get what's called an *Autoregressive Model* of order 2 (for the 2 lags) with a trend

$$y_t = 17 + 0.2t + 1.5y_{t-1} - 0.6y_{t-2} + \varepsilon_t.$$

The Secret Variable Revealed. At this point, Spanos revealed that x_t was the number of pairs of shoes owned by his grandmother over the observation period! She lives in the mountains of Cyprus, and at last count continues to add to her shoe collection. You will say this is a quirky made-up example, sure. It serves as a "canonical exemplar" for a type of erroneous inference. Some of the best known spurious correlations can be explained by trending means. For live exhibits, check out an entire website by Tyler Vigen devoted to exposing them. I don't know who collects statistics on the correlation between death by getting tangled in bed sheets and the consumption of cheese, but it's exposed as nonsense by the trending means. An example from philosophy that is similarly scotched is the case of sea levels in Venice and the price of bread in Britain (Sober 2001), as shown by Spanos (2010d, p. 366). In some cases, x is a variable that theory suggests is doing real work; discovering the misspecification effectively falsifies the theory from which the statistical model is derived.

I've omitted many of the tests, parametric and non-parametric, single assumption and joint (several assumptions), used in a full application of the same ideas, and mentioned only the bare graphs for simplicity. As you add questions you might wish to pose, they become your new primary inferences. The first primary statistical inference might indicate an effect of a certain magnitude passes with severity, and then background information might enter to tell if it's substantively important. At yet another level, the question might be to test a new model with variables to account for the trending mean of an earlier stage, but this gets beyond our planned M-S testing itinerary. That won't stop us from picking up souvenirs.

Souvenir V: Two More Points on M-S Tests and an Overview of Excursion 4

M-S Tests versus Model Selection: Fit Is Not Enough. M-S tests are distinct from model selection techniques that are so popular. Model selection begins with a family of models to be ranked by one or another criterion. Perhaps the most surprising implication of statistical inadequacy is to call into question the most widely used criterion of model selection: the goodness-of-fit/prediction measures. Such criteria rely on the "smallness" of the residuals. Mathematical fit isn't the same as what's needed for statistical inference. The residuals can be "small" while systematically different from white noise.

Members of the model selection tribe view the problem differently. Model selection techniques reflect the worry of overfitting: that if you add enough factors (e.g., $n - 1$ for sample size n), the fit can be made as good as desired, even if the model is inadequate for future prediction. (In our examples the factors took the form of trends or lags.) Thus, model selection techniques make you pay a penalty for the number of factors. We share this concern – it's too easy to attain fit without arriving at an adequate model. The trouble is that it remains easy to jump through the model selector's hoops, and still not achieve model adequacy, in the sense of adequately capturing the systematic information in the data. The goodness-of-fit measures already assume the likelihood function, when that's what the M-S tester is probing.

Take the Akaike Information Criterion (AIC) developed by Akaike in the 1970s (Akaike 1973). (There are updated versions, but nothing in our discussion depends on this.) The best known defenders of this account in philosophy are Elliott Sober and Malcolm Forster (Sober 2008, Forster and Sober 1994). An influential text in ecology is by Burnham and Anderson (2002). Model selection begins with a family of models such as the LRM: $y_t = \beta_0 + \beta_1 x_t + u_t$. They ask: Do you get a better fit – smaller residual – if you add $x^2{}_t$? What about adding both $x^2{}_t$ and $x^3{}_t$ terms? And so on. Each time you add a factor, the fit improves, but Akaike kicks you in the shins and handicaps you by 1 for the additional parameter. The result is a preference ranking of models by AIC score.[4] For the granny shoe data above, the model that AIC likes best is

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x^2{}_t + \beta_3 x^3{}_t + u_t.$$

Moreover, it's selection is within this family. But we know that all these LRMs are statistically inadequate! As with other purely comparative measures, there's no falsification of models.

What if we start with the adequate model that the PR arrived at, the autoregressive model with a trend? In that case, the AIC ranks at the very top of the model with the wrong number of trends. That is, it ranks a statistically inadequate model higher than the statistically adequate one. Moreover, the Akaike method for ranking isn't assured of having decent error probabilities. When the Akaike ranking is translated into a N-P test comparing this pair of models, the Type I error probability is around 0.18, and

[4] For each y_t form the residual squared. The sum of the squared residuals:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n}\left(y_t - \hat{\beta}_0 - \sum_{j=1}^{3}\hat{\beta}_j x_t^j\right)^2$$

gives an estimate of σ^2 for the model.

The AIC score for each contender in the case of the LRM, with sample size n, is $\log(\hat{\sigma}^2) + 2K/n$, where K is the number of parameters in model i. The models are then ranked with the smallest being preferred. The log-likelihood is the goodness-of-fit measure which is traded against simplicity, but if the statistical model is misspecified, one is using the wrong measure of fit.

For a comparison of the AIC using these data, and a number of related model-selection measures, see Spanos (2010a). None of these points change using the unbiased variant of AIC.

no warning of the laxity is given. As noted, model selection methods don't hone in on models outside the initial family. By contrast, building the model through our M-S approach is intended to accomplish both tasks – building and checking – in one fell swoop.

Leading proponents of AIC, Burnham and Anderson (2014, p. 627), are quite critical of error probabilities, declaring "P values are not proper evidence as they violate the likelihood principle (Royall 1997)." This tells us their own account forfeits control of error probabilities. "Burnham and Anderson (2002) in their textbook on likelihood methods for assessing models warn against data dredging . . . But there is nothing in the evidential measures recommended by Burnham and Anderson" to pick up on this (Dienes 2008, p. 144). See also Spanos (2014).

M-S Tests and Predesignation. Don't statistical M-S tests go against the error statistician's much-ballyhooed requirement that hypotheses be predesignated? The philosopher of science Rosenkrantz says yes:

> [O]rthodox tests . . . show how to test underlying assumptions of randomness, independence and stationarity, where none of these was the predesignated object of the test (the "tested hypothesis"). And yet, astoundingly in the face of all this, orthodox statisticians are one in their condemnation of "shopping for significance," picking out significant correlations in data post hoc, or "hunting for trends. . .". It is little wonder that orthodox tests tend to be highly ambivalent on the matter of predesignation. (Rosenkrantz 1977, 204–5)

Are we hoisted by our own petards? No. This is another case where failing to disentangle a rule's *raison d'être* leads to confusion. The aim of predesignation, as with the preference for novel data, is to avoid biasing selection effects in your primary statistical inference (see Tour III). The data are remodeled to ask a different question. Strictly speaking our model assumptions are predesignated as soon as we propose a given model for statistical inference. These are the pigeonholes in the PR menu. It has never been a matter of the time – of who knew what, when – but a matter of avoiding erroneous interpretations of the data at hand. M-S tests in the error statistical methodology are deliberately designed to be independent of (or orthogonal to) the primary question at hand. The model assumptions, singly or in groups, arise as argumentative assumptions, ready to be falsified by criticism. In many cases, the inference is as close to a deductive falsification as to be wished.

Parametric tests of assumptions may themselves have assumptions, which is why judicious combinations of varied tests are called upon to ensure their overall error probabilities. Order matters: Tests of the distribution, e.g., Normal, Binomial, or Poisson, assume IID, so one doesn't start there. The inference in the case of an M-S test of assumptions is not a statistical inference to a *generalization*: It's explaining given data, as with explaining a "known effect," only keeping to the statistical categories of distribution, independence/dependence, and homogeneity/heterogeneity (Section 4.6). Rosenkrantz's concerns pertain to the kind of pejorative hunting for variables to include in a substantive model. That's always kept distinct from the task of M-S testing, including respecifying.

Our argument for a respecified model is a *convergent* argument: questionable conjectures along the way don't bring down the tower (section 1.2). Instead, problems ramify so that the specification finally deemed adequate has been sufficiently severely tested for the task at hand. The trends and perhaps the lags that are required to render the statistical model adequate generally cry out for a substantive explanation. It may well be that different statistical models are adequate for probing different questions.[5] Violated assumptions are responsible for a good deal of non-replication, and yet it has gone largely unattended in current replication research.

Take-away of Excursion 4. For a severe tester, a crucial part of a statistical method's objectivity (Tour I) is registering how test specifications such as sample size (Tour II) and biasing selection effects (Tour III) alter its error-probing capacities. Testing assumptions (Tour IV) is also crucial to auditing. If a probabilist measure such as a Bayes factor is taken as a gold standard for critiquing error statistical tests, significance levels and other error probabilities appear to overstate evidence – at least on certain choices of priors. From the perspective of the severe tester, it can be just the reverse. Preregistered reports are promoted to advance replication by blocking selective reporting. Thus there is a tension between preregistration and probabilist accounts that downplay error probabilities, that declare them only relevant for long runs, or tantamount to considering hidden intentions. Moreover, in the interest of promoting Bayes factors, researchers who most deserve censure are thrown a handy life preserver. Violating the LP, using the sampling distribution for inferences with the data at hand, and the importance of error probabilities form an interconnected web of severe testing. They are necessary for every one of the requirements for objectivity.

[5] When two different models capture the data adequately, they are called *reparameterizations* of each other.