Diagnostics Cannot Have Much Power Against General Alternatives

David A. Freedman

Abstract

Model diagnostics are shown to have little power unless alternative hypotheses can be narrowly defined. For example, independence of observations cannot be tested against general forms of dependence. Thus, the basic assumptions in regression models cannot be inferred from the data. Equally, the proportionality assumption in proportional-hazards models is not testable. Specification error is a primary source of uncertainty in forecasting, and this uncertainty will be difficult to resolve without external calibration. Model-based causal inference is even more problematic.

Key words and phrases

Specification error, specification tests, model testing, forecast uncertainty, causal inference.

1. Introduction

The object here is to sketch a demonstration that, unless additional regularity conditions are imposed, model diagnostics have power only against a circumscribed class of alternative hypotheses. The paper is organized around the familiar requirements of statistical models. Theorems 1 and 2, for example, consider the hypothesis that distributions are continuous and have densities. According to the theorems, such hypotheses cannot be tested without additional structure.

Let us agree, then, that distributions are smooth. Can we test independence? Theorems 3 and 4 indicate the difficulty. Next, we grant independence and consider tests that distinguish between (i) independent and identically distributed random variables on the one hand, and (ii) independent but differently distributed variables on the other. Theorem 5 shows that, in general, power is lacking.

For ease of exposition, we present results for the unit interval; transformation to the positive half-line or the whole real line is easy. At the end of the paper, we specialize to more concrete situations, including regression and proportional-hazards models. We consider the implications for forecasting, mention some pertinent literature, and make some recommendations.

Definitions. A randomized test function is a measurable function ϕ with $0 \leq \phi(x) \leq 1$ for all x. A non-randomized test function ϕ has $\phi(x) = 0$ or 1. The size of ϕ is the supremum of $\int \phi \, d\mu$ over μ that satisfy the null hypothesis, a set of probabilities that will be specified in Theorems 1–5 below. The power of ϕ at a particular μ satisfying the alternative hypothesis is $\int \phi \, d\mu$. A simple hypothesis describes just one μ; otherwise, the hypothesis is composite. Write λ for Lebesgue measure on the Borel subsets of $[0, 1]$.

Interpretation. Given a test ϕ and data x, we reject the null with probability $\phi(x)$. Size is the maximal probability of rejection at μ that satisfy the null. Power at μ is the probability of rejection, defined for μ that satisfy the alternative.

Theorem 1. Consider probabilities μ on the Borel unit interval. Consider testing the simple null hypothesis

N: $\mu = \lambda$

against the composite alternative

A: μ *is a point mass at some (unspecified) point.*

Under these circumstances, any test of size α has power at most α against some alternatives.

Proof. Let ϕ be a randomized test function. If $\phi(x) > \alpha$ for all $x \in [0, 1]$, then $\int \phi(x) \, dx > \alpha$. We conclude that $\phi(x) \leq \alpha$ for some x, indeed, for a set of x's of positive Lebesgue measure. QED

Comments. (i) If we restrict ϕ to be non-randomized, then $\phi(x) = 0$ for some x. In other words, power would be 0 rather than α. (ii) The conclusions hold not just for some alternatives, but for many of them.

Theorem 2 requires some additional terminology. A "continuous" probability assigns measure 0 to each point. A "singular" probability on $[0, 1]$ concentrates on a set of Lebesgue measure 0.

Theorem 2. Consider probabilities μ on the Borel unit interval. Consider testing the simple null hypothesis

N: $\mu = \lambda$

against the composite alternative

A: μ *is continuous and singular.*

Under these circumstances, any test of size α has power at most α against some alternatives.

Proof. We identify 0 and 1, then visualize $[0, 1)$ as the additive group modulo 1 with convolution operator $*$. If μ is any probability, then $\lambda * \mu = \lambda$. Let ϕ be a randomized test function of size α. Then $\alpha \geq \int \phi \, d\lambda = \int\int \phi(x + y) \, \mu(dx) \, dy$. Hence, there are y with $\alpha \geq \int \phi(x + y) \, \mu(dx) = \int \phi(x) \, \mu_y(dx)$, where μ_y is the

translation of μ by y. If μ is continuous and singular, so is μ_y; but ϕ only has power α against μ_y. QED

Comments. (i) If we restrict ϕ to be non-randomized, then $\lambda\{\phi = 0\} \geq 1 - \alpha > 0$; the trivial case $\alpha = 1$ must be handled separately. Hence, power would be 0 rather than α. (ii) There are tests with high power against any particular alternative. Indeed, if ν is singular, it concentrates on a Borel set B with $\lambda(B) = 0$; let ϕ be the indicator function of B. This test has size 0, and power 1 at ν. The problem lies in distinguishing λ from the cloud of *all* alternatives.

A little more terminology may help. If μ and ν are two probabilities on the same σ-field, then μ is equivalent to ν if they have the same null sets. By the Radon-Nikodym theorem, this is tantamount to saying that the derivative of μ with respect to ν is positive and finite a.e.

Write λ^2 for Lebesgue measure on the Borel subsets of the unit square. Let ξ_1 and ξ_2 be the coordinate functions, so that $\xi_1(x, y) = x$ and $\xi_2(x, y) = y$. More generally, we write λ^k for Lebesgue measure on the Borel subsets of $[0, 1]^k$ and ξ_i for the coordinate functions, so $\xi_i(x_1, x_2, \ldots) = x_i$.

If μ is a probability on the unit square, let ρ_μ be the correlation between ξ_1 and ξ_2, computed according to μ. This is well-defined unless μ concentrates on a horizontal or vertical line.

For the proof of Theorem 3, if f is an integrable Borel function on the unit interval, then λ-almost all $x \in (0, 1)$ are Lebesgue points, in the sense that

$$\lim_{h \to 0} \frac{1}{h} \int_x^{x+h} f \, d\lambda \to f(x). \tag{1}$$

The result extends to k-dimensional space. See, for instance, Dunford and Schwartz (1958, p. 215).

Theorem 3. Consider probabilities μ on the Borel unit square that are equivalent to λ^2. Consider testing the simple null hypothesis

N: $\mu = \lambda^2$

against the composite alternative

A: $\rho_\mu \neq 0$.

Under these circumstances, a non-randomized test of size $\alpha < 1/2$ has power arbitrarily close to 0 at some alternative μ with $|\rho_\mu|$ arbitrarily close to 1; furthermore, μ can be chosen to have a continuous positive density.

Proof. Consider a non-randomized test function ϕ with size α. Let G be the subset of the unit square where $\phi = 0$. So $\lambda^2(G) \geq 1 - \alpha > 1/2$. Let G^* be the set of pairs $(1 - x, 1 - y)$ with $(x, y) \in G$. So $\lambda^2(G^*) > 1/2$ and $\lambda^2(G \cap G^*) > 0$. We can find u, v with $u \neq 1/2$, $v \neq 1/2$ and (u, v) a Lebesgue point of $G \cap G^*$.

Thus, (u, v) is a Lebesgue point of G, and so is $(1 - u, 1 - v)$. These two points are different, and lie on a line of non-zero slope; if we put mass $1/2$ at each point, the correlation between ξ_1 and ξ_2 would be ± 1.

Now construct a continuous positive density f that puts mass nearly $1/2$ in a small neighborhood of (u, v), and mass nearly $1/2$ in a small neighborhood of $(1 - u, 1 - v)$. With respect to f, the correlation between ξ_1 and ξ_2 is essentially ± 1. Moreover, $\int_G f \, d\lambda^2$ is nearly 1, so $\int \phi f \, d\lambda^2$ is nearly 0. QED

Comment. The correlation is used to pick out alternatives that are quite different from independence.

Let $\mathcal{D}$ be the set of densities on $[0, 1]^k$ that can be represented as a finite sum $\sum_j c_j g_j$, where $c_j \geq 0$, $\sum_j c_j = 1$, $x = (x_1, \ldots, x_k)$, and

$$g_j(x) = \prod_{i=1}^{k} g_{ij}(x_i), \tag{2}$$

the g_{ij} being continuous densities on $[0, 1]$. Unless otherwise specified, densities are with respect to Lebesgue measure.

Lemma 1. The uniform closure of $\mathcal{D}$ is the set of all continuous densities on $[0, 1]^k$.

Proof. This is easily done, using k-dimensional Bernstein polynomials. See Lorentz (1986, p. 51). QED

Lemma 2. Let ϕ be a randomized test function on $[0, 1]^k$. Suppose $\int \phi \, d\mu \leq \alpha$ for all probabilities μ on $[0, 1]^k$ that make the coordinate functions independent with continuous positive densities. Then ess. sup. $\phi \leq \alpha$.

Proof. The condition is

$$\int_{[0,1]^k} \phi \times \left(\prod_{i=1}^{k} f_i \right) d\lambda^k \leq \alpha \tag{3}$$

for all continuous positive densities f_i on $[0, 1]$. By an easy passage to the limit, inequality (3) holds for all continuous densities; that is, $f_i \geq 0$ rather than $f_i > 0$.

A convexity argument shows that

$$\int \phi \varphi \, d\lambda^k \leq \alpha \tag{4}$$

for any $\varphi \in \mathcal{D}$, hence for any continuous density φ on $[0, 1]^k$ by Lemma 1. A density on $[0, 1]^k$ can be approximated in L^1 by a continuous density. Thus, inequality (4)

holds for all densities φ on $[0, 1]^k$. Let $B = \{\phi > \alpha\}$. Suppose $\lambda^k(B) > 0$. Let $\varphi = 1/\lambda^k(B)$ on B, and let φ vanish off B. Then $\alpha \geq \int \phi\varphi \, d\lambda^k \geq \int_B \phi\varphi \, d\lambda^k > \alpha \int_B \varphi \, d\lambda^k = \alpha$, a contradiction showing that $\lambda^k(B) = 0$. QED

Theorem 4. Consider probabilities μ on the Borel unit hypercube $[0, 1]^k$ that are equivalent to λ^k. Consider testing the composite null hypothesis

> N: *the coordinate functions are independent with respect to μ, each coordinate having a continuous positive density*

against any alternative set A of μ's. Under these circumstances, any test of size α has power at most α.

Proof. This is immediate from Lemma 2. QED

Comment. The alternative A can consist of a single density f that is positive a.e. but is otherwise badly behaved. The null hypothesis can be substantially restricted, say to polynomial densities.

Theorem 5 is couched in terms of remote alternatives, which are distant from the null hypothesis. For rigor, we would have to metrize the space of probabilities on $[0, 1]^k$. This can be done in several ways without changing the argument. Here are three possibilities. (i) Variation distance can be used; remote alternatives will be nearly singular with respect to the probability satisfying the (simple) null hypothesis. (ii) The sup norm on distribution functions is another possibility; remote alternatives will be at a distance nearly 1 from the null. (iii) Distance can also be defined so as to metrize the weak-star topology; remote alternatives will be at a substantial distance from the null, with details depending a little on the metric that is used. Recall that λ^k is Lebesgue measure on the Borel subsets of $[0, 1]^k$.

Theorem 5. Fix α with $0 < \alpha < 1$. Fix δ with $\delta > 0$ and $\alpha + \delta < 1$. Let μ be a probability on $[0, 1]^k$, and let $\xi_1, \ldots, \xi_k$ be the coordinate functions. Consider testing the simple null hypothesis

> N: $\mu = \lambda^k$

against the composite alternative

> A: *μ makes the coordinate functions independent with different distributions, each distribution having a continuous positive density on $[0, 1]$.*

There is a fixed positive integer k_0 such for any $k > k_0$, and any randomized test ϕ_k of size α, there is a remote alternative μ where power is less than $\alpha + \delta$. The alternative μ makes ξ_i independent with continuous positive density f_i, but each f_i is arbitrarily concentrated near some point c_i in $(0, 1)$. Moreover,

> (i) $c_i < 1/3$ *for more than $k/4$ of the indices $i = 1, \ldots, k$, and*
> (ii) $c_i > 2/3$ *for more than $k/4$ of the indices $i = 1, \ldots, k$.*

Proof. *Step 1.* Let ϕ_k be a randomized test function of size α on $[0, 1]^k$, so

$$\int \phi_k \, d\lambda^k \leq \alpha. \tag{5}$$

Apply Markov's inequality to $1 - \phi_k$, to see that for all k,

$$P(B_k) \geq \frac{\delta}{\alpha + \delta}, \quad \text{where } B_k = \{\phi_k < \alpha + \delta\}. \tag{6}$$

Step 2. Write 1_Q for the indicator function of the set Q. Let C_k be the subset of $[0, 1]^k$ where $\xi_i < 1/3$ for more than $k/4$ of the indices $i = 1, \ldots, k$. Formally,

$$C_k = \left\{ \sum_{i=1}^{k} 1_{\{\xi_i < 1/3\}} > k/4 \right\}. \tag{7}$$

Similarly, let

$$D_k = \left\{ \sum_{i=1}^{k} 1_{\{\xi_i > 2/3\}} > k/4 \right\}. \tag{8}$$

The ξ_i are independent and $\lambda^k\{\xi_i < 1/3\} = 1/3 > 1/4$. So $\lambda^k(C_k) \to 1$ by the law of large numbers, and likewise for D_k.

Step 3. Choose k_0 so that

$$\lambda^k(C_k \cap D_k) \geq 1 - \frac{1}{2} \frac{\delta}{\alpha + \delta} \quad \text{for } k > k_0. \tag{9}$$

Then

$$\lambda^k(B_k \cap C_k \cap D_k) \geq \frac{1}{2} \frac{\delta}{\alpha + \delta} \quad \text{for } k > k_0. \tag{10}$$

Step 4. There must be $c_i \in (0, 1)$ with $c = (c_1, \ldots, c_k)$ a Lebesgue point of $B_k \cap C_k \cap D_k$. For each i, we construct a continuous positive density f_i on $[0, 1]$ that is concentrated around c_i, with

$$\int_{[0,1]^k} \phi_k \times \left(\prod_{i=1}^{k} f_i \right) d\lambda^k \leq \phi_k(c) + \delta \leq \alpha + \delta. \tag{11}$$

To get the densities, start by putting the uniform distribution on a small hypercube centered at $(c_1, \ldots, c_k)$. Then smooth each edge separately. QED

Comments

(i) If the test function is not randomized, we get power approaching 0 at remote alternatives, by the same argument.

(ii)   The alternatives are remote from the null λ^k. They are also radically different from any power measure θ^k.

(iii) What if we have a suite of diagnostics? Let ϕ be the final result: 1 if the model is rejected, 0 if accepted, and $0 < \phi < 1$ if results are ambiguous or the decision is randomized. Thus, it suffices to consider a single test function.

(iv)   Lehmann and Romano (2005, §14.6) discuss statistical literature on limits to diagnostic power, the chief result being due to Janssen (2000). Even with IID data and a simple null hypothesis, goodness-of-fit tests have little power unless the set of alternative hypotheses can be substantially restricted. The context, however, is "local alternatives" that are const.$/\sqrt{k}$ away from the null; such alternatives tend to the null as the sample size k increases. There are also some relevant papers in the game-theory literature: for instance, Lehrer (2001), Olszewski and Sandroni (2008).

(v)   By way of comparison, with IID data, the Kolmogorov-Smirnov test (among others) has power approaching 1 at any fixed alternative when testing the simple null hypothesis that the common distribution is uniform on [0, 1]. In Theorem 5 above, the data are IID under the null not the alternative, and the sample size k is fixed.

2. Specific models

In regression models, a key assumption is exogeneity: variables included in the model should be independent of error terms. Theorems 3 and 4 indicate the difficulty of testing this assumption. For many purposes, errors may be required to be independent, or independent and identically distributed. Theorems 3–5 indicate the difficulties. Requiring orthogonality rather than independence will not solve the problem, and may not suffice for the usual asymptotics.

Rather than pursuing such topics, we turn to the proportional-hazards model, where subjects have failure times and censoring times. These are positive random variables, and only the smaller of the two is observed. These variables are generally assumed to be independent, or conditionally independent in a suitable sense given certain information. For a review of the model, see Freedman (2008).

As shown by Tsiatis (1975), independence of failure times and censoring times is not testable in the usual data structures. Without that assumption, marginal distributions are not identifiable. Also see Clifford (1977). Furthermore, what happens after the end of a study is plainly unknowable. This already puts severe limits on the power of diagnostics. Therefore, let us assume that all failure times are fully observable, and see what can done in that context.

In the model, failure times τ_i are independent positive variables with absolutely continuous distribution functions F_i. The density is $f_i = F_i{}'$ and the hazard rate is $h_i = f_i/(1 - F_i)$. According to the model, there is a baseline hazard rate h, and $h_i = h \exp(X_i \beta)$, where β is a parameter, or parameter vector if X_i is a vector. For

present purposes, the covariate X_i is allowed to depend on i but not on time, and X_i is non-stochastic; we require $0 < h_i < \infty$ a.e. More general forms of the model relax these restrictions.

Theorems 1 and 2 show that we cannot tell whether failure times have densities and hazard rates—unless we restrict the class of alternatives, or impose additional assumptions, qualifications that will not be repeated. Theorems 3 and 4 shows that we cannot determine whether failure times are independent or dependent.

Let us therefore assume that failure times have continuous positive densities. Even so, Theorem 5 shows that we cannot test the proportional-hazards assumption; this takes some mathematical effort to verify. Let us begin with the null hypothesis that the baseline hazard rate is identically 1, corresponding to a standard exponential failure time, and all the covariates are identically 0.

We can reduce to the IID uniform case covered by the theorem. All it takes is a change of variables: replace the failure time τ_i by $\exp(-\tau_i)$. The conclusion is that any test of size α will have power barely above α against certain remote alternatives; the latter make the τ_i independent with continuous positive densities on $(0, \infty)$, but highly concentrated.

More general null hypotheses follow the same pattern. Suppose the covariates X_i are linearly independent p-vectors. The parameter vector β is a p-vector too. The null hypothesis specifies that the baseline hazard rate is 1, but allows the ith subject to have the hazard rate $\exp(X_i\beta)$, where β is free to vary—although it must be constant across i's.

We can again replace τ_i by $\exp(-\tau_i)$. Theorem 5 will give independent τ_i with highly concentrated densities at which power is low. Over 1/4 of these densities will be concentrated at values larger than $\log 3$, and will be quite different from any exponential density.

Comments

 (i) According to the theorem, for each large k, there are remote alternatives that are nearly indistinguishable from the null.

(ii) Replacing the baseline hazard rate by an unknown h does not change the position.

(iii) Usual tests of the proportional-hazards model involve adding another covariate, or stratifying on that covariate; the implied alternatives are tamer than the ones constructed here.

(iv) Altman and de Stavola (1994) discuss some of the practical problems in testing proportional-hazards models, and note that power is generally limited—even with conventional alternatives.

Theorem 6. Suppose $\tau_i : i = 1, \ldots, k$ are positive random variables, each having a distribution with a continuous positive density on $(0, \infty)$. Suppose the τ_i are independent. Under these circumstances, the τ's obey a proportional-hazards model

with a pre-specified baseline hazard rate h_0. The covariates are non-stochastic but time-dependent.

Proof. We construct the model as follows. Let h_i be the hazard rate of τ_i. The covariate are $X_{it} = \log h_i(t) - \log h_0(t)$ for $i = 1, 2, 3, \ldots$. Furthermore, $h_i(t) = h_0(t) \exp(X_{it}\beta)$ with $\beta = 1$. QED

Comment. Unless we restrict the set of covariates, the proportional-hazards model includes all distributions for failure times.

It would seem that Theorem 5 can be modified to handle time-varying non-stochastic covariates. But that includes all distributions, according to Theorem 6. To resolve the air of paradox, let ϕ be a randomized test function on $[0, 1]^k$ with

$$\int_{[0,1]^k} \phi \times \left(\textstyle\prod_{i=1}^k f_i \right) d\lambda^k \leq \alpha \qquad (12)$$

for all continuous, positive densities f_i. By Lemma 2, ess. sup. $\phi \leq \alpha$. There is no contradiction, because there are no non-trivial tests.

3. Discussion

A Google search (performed on March 18, 2008) gave 53,000 hits on the phrase "regression diagnostics," so this is a topic of some interest. Amazon.com gave 1300 hits on the phrase, the two most relevant books being Belsley et al. (2004) and Fox (1991). These texts do not reach the issues discussed here. Diagnostics for the proportional hazards model are frequently mentioned, but standard references do not indicate the limitations on power; see, for instance, Andersen and Keiding (2006).

Models are frequently used to make causal inferences from observational data. See *Scandinavian Journal of Statistics* 2004, vol. 31, no. 2, for a recent survey. In brief,

> "Fortunately, the days of 'statistics can only tell us about association, and association is not causation' seem to be permanently over." [p. 161]

For causal inference, the crucial assumption is "invariance to intervention:" statistical relationships, including parameter values, that obtain in an observational setting will also obtain under intervention. For discussion and some historical background, see Freedman (2005).

The invariance assumption is not entirely statistical. Absent special circumstances, it does not appear that the assumption can be tested with the data that are used to fit the model. Indeed, it may be difficult to test the assumption without an experiment, either planned or natural. Such tests are beyond the scope of this paper. They are also beyond the scope of conventional diagnostic procedures.

4. What about forecasting?

In principle, forecasting should be easier than making causal inferences from observational data, because forecasts are more readily calibrated against outcomes. On the other hand, the system that we are forecasting may be unstable, or we may be interested in forecasting rare events, or we may need the forecast before calibration data are available.

Conventional models seem to offer abundant ways to measure forecast uncertainty, just based on the data at hand. We can compute R^2 or the standard error of regression (40,000 hits on the latter phrase in Google); we can use cross-validation (1,000,000 hits) and so forth. However, as recent economic history makes clear, a major source of uncertainty in forecasts is specification error in the forecasting models. Specification error is extremely difficult to evaluate using internal evidence. That is the message of the present paper.

Standard econometric texts, like Greene (2007) or Kennedy (2003), spend many pages discussing specification error, regression diagnostics (aka specification tests, model checking), robust estimation, and similar topics. Caution is in order. Unless the relevant class of specification errors can be narrowly delimited by prior theory and experience, diagnostics have limited power, and the robust procedures may be robust only against irrelevant departures from assumptions. "Robust standard errors" are particularly misleading, since these ignore bias (Freedman, 2006).

5. Recommendations

Model diagnostics are seldom reported in applied papers. My recommendation, which may seem paradoxical at first, is this. Diagnostics should be reported more often, but a skeptical attitude should be adopted toward the results. Diagnostics should be reported more often because they can yield helpful information, picking up specification errors if these are sufficiently gross. Such errors might be corrected by adding explanatory variables, or modifying assumptions about disturbance terms, or changing the functional form of the equation. Furthermore, greater transparency in model development would eventually make the whole enterprise more credible. On the other hand, skepticism about diagnostics is warranted. As shown by the theorems presented here, a model can pass diagnostics with flying colors yet be ill-suited for the task at hand.

References

Altman, D. G., & de Stavola, B. L. (1994). Practical problems in fitting a proportional hazards model to data with updated measurements of the covariates. *Statistics in Medicine*, *13*, 301–41.

Andersen, P. K., & Keiding, N., eds. (2006). *Survival and Event History Analysis*. Chichester, U. K.: John Wiley & Sons.

Belsley, D. A., Kuh, E., & Welsch, R. E. (2004). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley-Interscience.

Clifford, P. (1977). Nonidentifiability in stochastic models of illness and death. *Proceedings of the National Academy of Sciences*, *USA*, *74*, 1338–40.

Doob, J. L. (1953). *Stochastic Processes*. New York: Wiley.

Dunford, N., & Schwartz, J. T. (1958). *Linear Operators: Part I*, *General Theory*. New York: Interscience.

Fox, J. (1991). *Regression Diagnostics: An Introduction*. Thousand Oaks, CA: Sage Publications, Inc.

Freedman, D. A. (2008). Survival analysis: A primer. *The American Statistician*, *62*, 110–119.

Freedman, D. A. (2006). On the so-called "Huber Sandwich Estimator" and "robust standard errors." *The American Statistician*, *60*, 299–302.

Freedman, D. A. (2005). *Statistical Models: Theory and Practice*. New York: Cambridge University Press.

Greene, W. H. (2007). *Econometric Analysis*, 6th ed. Prentice Hall. NJ: Upper Saddle River.

Janssen, A. (2000). Global power functions of goodness-of-fit tests. *Annals of Statistics*, *28*, 239–53.

Kennedy, P. (2003). *A Guide to Econometrics*, 5th ed. Cambridge, MA: MIT press.

Lehmann, E., & Romano, J. (2005). *Testing Statistical Hypotheses*, 3rd ed. New York: Springer.

Lehrer, E. (2001). Any inspection rule is manipulable. *Econometrica*, *69*, 1333–47.

Lorentz, G. G. (1986). *Bernstein Polynomials*, 2nd ed. New York: Chelsea.

Olszewski, W. and Sandroni, A. (2008). Manipulability of future-independent tests. *Econometrica*, to appear.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, *USA*, *72*, 20–22.

About the author. David A. Freedman is professor of statistics, University of California, Berkeley CA 94720-3860, USA. E-mail: freedman@stat.berkeley.edu