Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
International Journal of Forecasting 25 (2009) 833–839
international journal of forecasting
www.elsevier.com/locate/ijforecast

Diagnostics cannot have much power against general alternatives

David A. Freedman[1]

Department of Statistics, University of California, Berkeley, United States

Abstract

Model diagnostics are shown to have little power unless alternative hypotheses can be narrowly defined. For example, the 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.

Keywords: Specification error; Specification tests; Model testing; Forecast uncertainty; Causal inference

1. Introduction

The objective here is to demonstrate 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 $0 \le \phi(x) \le 1$ for all x. A non-randomized test function ϕ has $\phi(x) = 0$ or 1. The

[1] Professor David Freedman wrote this article for the *International Journal of Forecasting* shortly before his death in October 2008.

0169-2070/$ - see front matter © 2009 Published by Elsevier B.V. on behalf of International Institute of Forecasters.
doi:10.1016/j.ijforecast.2009.05.004

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 $\phi(x)$. The size is the maximal probability of rejection at μ that satisfies the null. The power at μ is the probability of rejection, defined for μ, that satisfies 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 alternative.

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 $\phi(x) \le \alpha$ for some x, or indeed, for a set of xs of positive Lebesgue measure. □

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 alternative.

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 $\alpha \ge \int \phi \, d\lambda = \int \int \phi(x + y) \, \mu(dx) \, dy$. Hence, there are y with $\alpha \ge \int \phi(x + y) \, \mu(dx) = \int \phi(x) \, \mu_y(dx)$, 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. □

Comments

(i) If we restrict ϕ to be non-randomized, then $\lambda\{\phi = 0\} \ge 1 - \alpha > 0$; 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

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

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