"FIND YOUR KEYS YET?" (SOME THOUGHTS ON PARAMETRIC MODEL MISSPECIFICATION)

JEFFREY S. RACINE

Statistical Framework

Consider a stochastic relationship involving a predictor $X \in \mathbb{R}^1$ and a response $Y \in \mathbb{R}^1$. Suppose interest lies in the average value of Y given $X = x$ which we denote $g(x)$ for a location-scale model given by $Y = g(X) + \sigma(X)\epsilon$. The (unknown) 'conditional mean' evaluated at $X = x$ is defined as

$$\begin{aligned} g(x) \equiv E(Y|X = x) &= \int_{-\infty}^{\infty} y \frac{f(y,x)}{f(x)}\, dy \\ &= \frac{\int_{-\infty}^{\infty} y f(y,x)\, dy}{f(x)} \\ &= \frac{m(x)}{f(x)}. \end{aligned}$$

The (unknown) response $\beta(x)$ is defined as

$$\begin{aligned} \beta(x) \equiv \frac{d\, g(x)}{dx} &= \frac{f(x)m'(x) - m(x)f'(x)}{f^2(x)} \\ &= \frac{m'(x)}{f(x)} - \frac{m(x)}{f(x)}\frac{f'(x)}{f(x)} \\ &= \frac{m'(x)}{f(x)} - g(x)\frac{f'(x)}{f(x)}. \end{aligned}$$

This model must satisfy the restrictions a) $E|Y| < \infty$, b) $f(x) = \int_{-\infty}^{\infty} f(y,x)\, dy$, $\int_{-\infty}^{\infty} f(x)\, dx = 1$, $f(x) \geq 0$, and c) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(y,x)\, dy dx = 1$, $f(y,x) \geq 0$ (i.e. the conditional mean exists, and $f(x)$ and $f(y,x)$ are proper density functions). Note that $f(y,x)$, $f(x)$, $g(x)$ and $\beta(x)$ are unknown functions of x (i.e. *functions* of x, not *constants*).

Date: September 12, 2011.

"The light is better here…"

Suppose we wish to model $g(x)$ and $\beta(x)$ in some sound way within a statistical framework (i.e. conduct 'regression analysis'). To begin with the statement "Assume that $\beta(x) \equiv \frac{m'(x)}{f(x)} - g(x)\frac{f'(x)}{f(x)}$ equals a constant" and then restrict one's attention to models possessing this property has always puzzled me. Yet the popular linear-in-variables parametric specification $g(x) = \beta_0 + \beta_1 x$ that is frequently used by practitioners possesses exactly this property since $d\,g(x)/dx = \beta_1$.[1]

In the absence of any reason whatsoever as to why this ought to be the case, this strikes me as sadly similar to the story of the person who lost their keys at night and is looking for them under a lamppost, and when approached by a stranger who asks "what are you doing?" and is given the reply "looking for my lost keys…" which is met by "where did you lose them?" and is given the reply "in the alley behind that building…" which is met by "why are you looking here then?" and the reply is "because the light is better here!" In the context of model specification I detect an unsettling similarity between behaviour that suggests someone would rather interpret scalars than vectors (i.e. the constant β_1 rather than $\beta(x)$) and behaviour that would lead an individual to look for their keys in some place solely because 'the light is better here'.[2]

Some perspective on parametric model specification might be in order. Often students are given a set of conditions under which parametric estimators possess certain desirable properties, and the first one is typically that the model one is using is 'correct' (i.e. the 'true' model). So this places the burden of locating the correct model upon the researchers' shoulder which requires writing down the 'correct' model. To paraphrase Aman Ullah, finding the correct model, like achieving nirvana, is known not to be an easy

[1] Not to mention all cross-partials being zero for the ever-popular simple multivariate linear model $g(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$, a consequence of additivity/strict separability etc., which are exceptionally strong presumptions to my way of thinking.

[2] I harbour the same disdain for semiparametric/nonparametric average derivative estimation (as if the average derivative having a $\sqrt{n}$ rate of convergence and being a scalar in any way constitutes some sort of compensating virtue). A simple example will suffice. Let $y = x^2$ and let x be uniformly distributed on $[-1, 1]$. The true derivative is $\beta(x) = 2x$ which is non-zero almost everywhere, negative for $x < 0$, positive otherwise, yet the average derivative is exactly zero. What can we possibly infer about the true derivative $\beta(x) = 2x$ based upon the average derivative? Not much to my way of thinking. Yet practitioners will then use the semiparametric/nonparametric average derivative value along with its standard error to compute the analog of a test of significance in a simple linear parametric model, the logic of which makes me cry.

task. When one pulls a function from the space of functions (a 'Banach Space' which is naturally dense) then pretends that they have the 'correct model' (i.e. the 'true data generating process' which is of course required for valid parametric inference), realize that the probability of drawing the 'correct model' from the space of functions is the same as the probability of drawing any *specific* realization of a random variable having a continuous distribution - even my first year statistics students appreciate this three weeks into the course and realize this is a 'measure-zero' event.

So, naturally, when I encounter simple linear models the first question that comes to mind is "did the model pass a specification test?" and I am often saddened to learn that no such testing was undertaken. This is not to say that such models may not be useful approximations, rather just to point out that properties that depend crucially on correct specification would not be expected to hold (such as presumed null distributions and critical values for all tests, for instance).

One can grow tired of such exercises (or for that matter sitting in seminars listening to someone launch right into the fascinating intuitive/counter-intuitive interpretation they ascribe to the constant $\hat{\beta}_1$ in their [misspecified almost surely] model which was not subjected to any specification testing whatsoever), just as one would grow tired of posing the question "Find your keys yet?" which sadly often has the same answer as the question "Did we actually learn anything here?"

Find your keys yet?