NONPARAMETRIC REGRESSION WHEN THE PREDICTOR IS SUBJECT TO ERROR

Work done jointly with

Jeff Maca, Novartis
R. J. Carroll, U. Penn.

OUTLINE

- Statement of problem
- Globally consistent nonparametric regression (Fan and Truong)
  - does not work so well
    - Fan & Truong show very poor asymptotic rates of convergence
    - we have simulations showing poor finite-sample behavior
- Standard measurement error method: SIMEX (functional)
  - Definition by graphs
- Alternative to SIMEX (structural)
  - Regression splines for basic regression model
  - Mixtures of normals for covariate density model
  - Emphasis is on flexible parametric modeling, not nonparametric modeling.
- Some simulations

THE PROBLEM OF MEASUREMENT ERROR

- We are interested in nonparametric regression when the predictor X cannot be observed exactly.
  - $Y = m(X) + \epsilon$ where m is only known to be smooth
  - Observe Y and $W = X + U$, where
    - $E(U) = 0$
    - $\mathrm{var}(U) = \sigma_u^2$
    - U normally distributed
  - The normality is not important.
- For today, measurement error variance σ_u^2 assumed known.
- Measurement error occurs in a wide variety of problems.
  - Measuring nutrient intake
  - Measuring airborne lead exposure
  - Measuring blood pressure
- The effects of measurement error are:
  - biased estimates of the regression line
  - increase the perceived variability about the regression line.
- Other than the work of Fan and Truong, there has been little done on nonparametric regression with measurement error.

EFFECTS OF MEASUREMENT ERROR

$$E(Y|X) = m(X), \quad W = X + U$$

- We know that if we do nonparametric regression and ignore the measurement error, then the estimate of $m(x)$ is badly biased.
- A sad fact: in our setup, it is known that the best that one can do with consistent estimation of $m(x)$ is a rate of $\{\log(n)\}^{-2}$.
  - $\log(16,000,000) \approx 10$.
  - $n^{-1/2}$ is the parametric error rate
  - $n^{-2/5}$ is the error rate for "standard" nonparametric problems
  - see Carroll–Hall–Stefanski for theory for density estimation, Fan–Truong for this problem.
  - Small global bias $\Rightarrow$ enormous variance.
- In other words, *globally consistent* nonparametric estimation in the presence of measurement error with reasonable precision is not possible.

WHAT TO DO?

- *Globally consistent* nonparametric estimation in the presence of measurement error with reasonable precision is not possible.
  - Simulations bear this out.
  - You get either lots of bias or lots of variance.
- Option #1: the problem is impossible, hence not worth further study.
- Option #2: relax from the term *globally consistent*.
- The intent of the talk is to describe a couple of methods which are either
  - approximately consistent globally or
  - consistent for a wide subclass of models.

SIMEX

- The SIMEX method is due to Cook & Stefanski (1995, JASA).
  - The theory is in Carroll, et al. (1996, JASA)
- SIMEX has been previously applied to parametric problems.
  - It makes no assumptions about the true X's. (*Functional*)
  - It results in estimators which are *approximately* consistent, i.e., consistent at least to order $O(\sigma_u^6)$.
- Here is the method, defined via graphs.

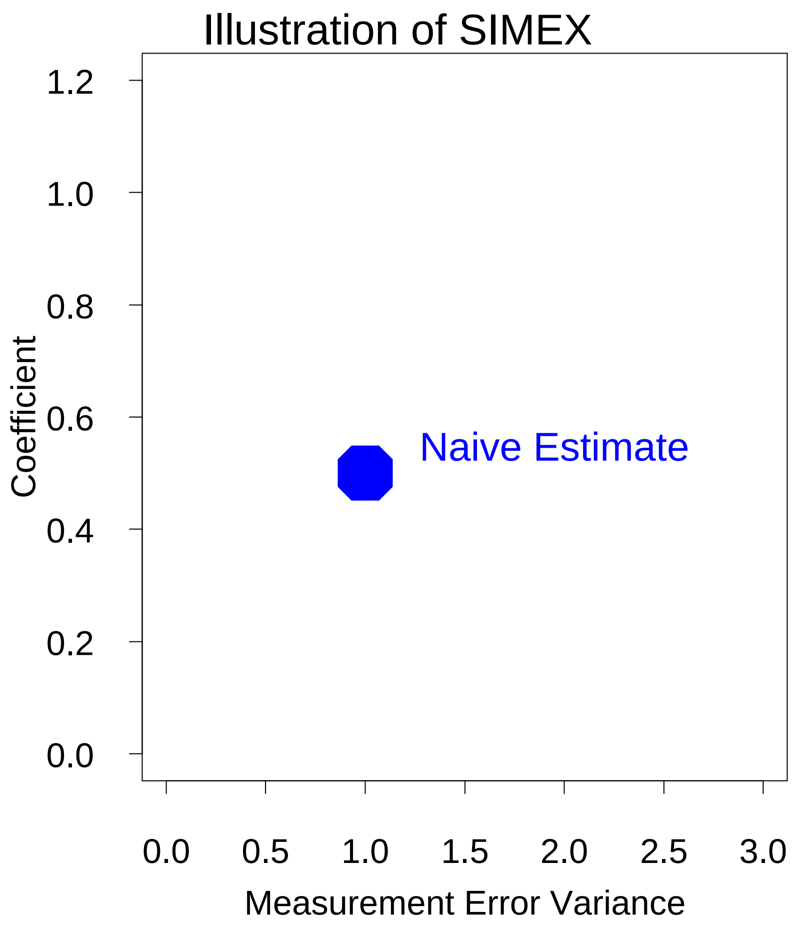
Illustration of SIMEX
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Coefficient
Naive Estimate
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Measurement Error Variance

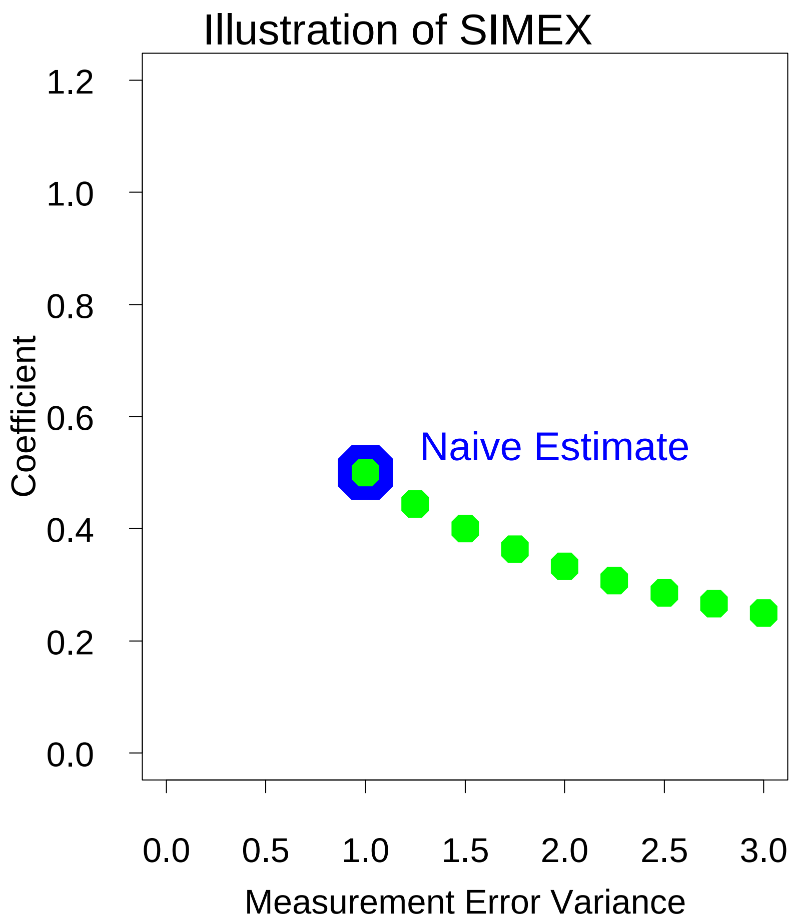
Illustration of SIMEX
Naive Estimate
Coefficient
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Measurement Error Variance

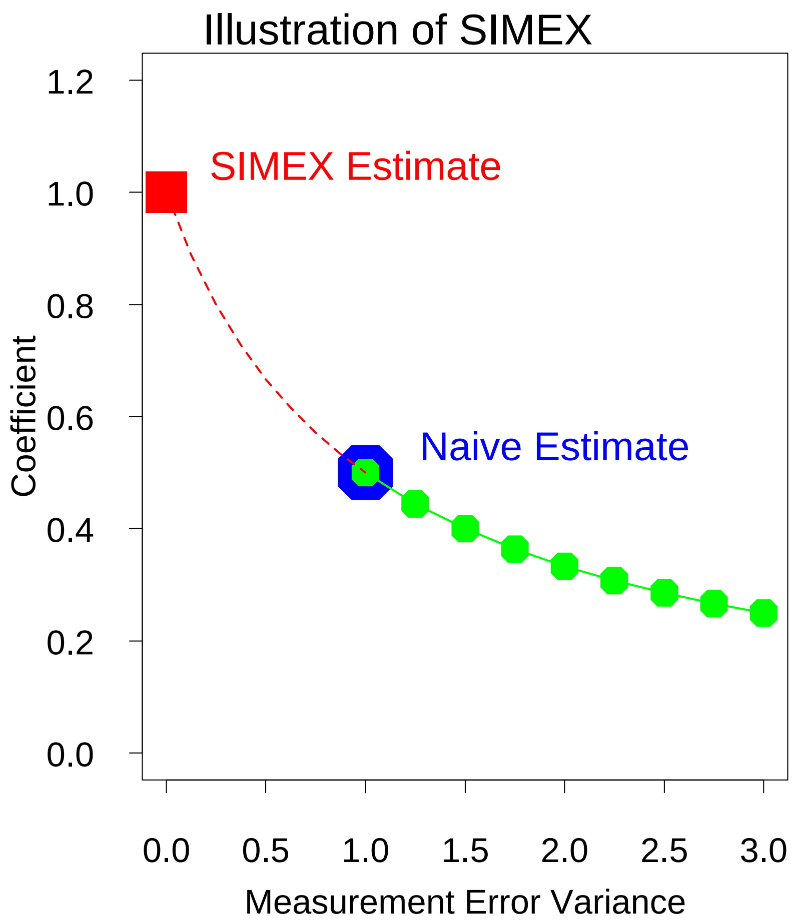
Illustration of SIMEX
SIMEX Estimate
Naive Estimate
Coefficient
Measurement Error Variance
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.0
0.5
1.0
1.5
2.0
2.5
3.0

SIMEX

- Our suggestion is to apply the SIMEX idea to nonparametric regression.
- We have a fully worked–out asymptotic theory in the kernel regression context.
  - The estimators have the usual rates of convergence.
  - They are approximately consistent, to order $O(\sigma_u^6)$.
  - In the SIMEX implementation with a grid of errors 1.5, 2.0, 2.5 and 3.0 times as large as the original error variance:
    - quadratic extrapolant is 9.0 times more variable than linear extrapolant
    - cubic extrapolant is 52.0 times more variable than linear extrapolant
    - These results are <u>independent</u> of the problem.
- An asymptotic theory with rates seems very difficult for splines
  - but, simulations indicate that SIMEX/splines works a little *better* than SIMEX/kernel

STRUCTURAL MODELING

- The basic idea is based upon the following notion:
- We can estimate $E(Y|W)$ from our data
- But, this is related to the desired function $m(X)$ by

$$E(Y|W) = E\{m(X)|W\} = \int m(x)f(x|W)dx.$$

- If we had a convenient form for $m(X)$, and if we knew the distribution of X given W, then we could estimate $m(X)$ by minimizing over the data

$$\sum_{i=1}^{n} \left\{Y_i - \int m(x)f(x|W_i)dx\right\}^2.$$

- We need two things to make this work:
  - convenient flexible form for $m(x)$;
  - convenient flexible distribution for X.

WHAT IF WE HAVE A DISTRIBUTION FOR X?

- The SIMEX method makes no assumptions about the distribution of X.
  - This is a striking robustness property.
- Suppose however that we knew the distribution of X, and that U is normally distributed.
  - We in particular know the distribution of X given W!
  - Could we get anywhere?
- Consider regression splines of order J with K knots:

$$E(Y|X) = \sum_{j=0}^{J} \beta_j X^j + \sum_{j=1}^{K} \beta_{j+J}(X - \xi_j)_+^J$$

- The terms $\xi_1, ..., \xi_K$ are knots, to build flexibility

REGRESSION SPLINES

$$E(Y|X) = \sum_{j=0}^{J} \beta_j X^j + \sum_{j=1}^{K} \beta_{j+J} (X - \xi_j)_+^J$$

- If we know the joint distribution of X and U, then the observed data have

$$E(Y|W) = \sum_{j=0}^{J} \beta_j E(X^j|W) + \sum_{j=1}^{K} \beta_{j+J} E\{(X - \xi_j)_+^J|W\}$$

- This is just a linear model in the β's!!!!
- There are many methods to fit such splines
  - Fixed knots, with shrinkage (Eilers and Marx, Ruppert and Carroll)
  - Knots selection and deletion (Friedman, Stone and coworkers)
- The key remaining issue: the joint distribution of X and U.

FLEXIBLE DISTRIBUTIONAL MODELING

- We have seen that we need a model for X and U
  - There is usually good information about the measurement error U from other studies, replication, etc. We will assume it is normal here (computationally convenient!)
- We thus need a flexible model for the distribution of X.
  - We use the model of mixtures of normals with an unknown number of mixtures.
  - Even more convenient, there have been a lot of recent papers on fitting mixtures of normals.
- Bayesian methods: Roeder & Wasserman (JASA, 1997) and Green & Richardson (1996, JRSS-B).
- We used Gibbs sampling to estimate the parameters Θ_L of the mixture of normals with L components.

THE ALGORITHMS

- The basic spline fitting method is penalized least-squares.
- Ridge smoothing via AIC, GCV, CV, etc. performed very badly because of occasional gross under–smoothing
- We developed a different smoothing based on estimated MSE, borrowing ideas from the kernel smoothing literature on plug–in estimation of bias.
- The SIMEX method was implemented pointwise.
  - The penalized least-squares method was used directly.
  - The fitted values (and not the regression coefficients) were extrapolated.
- The mixture of normals was fit using Gibbs sampling.
  - This yielded the distribution of X given W. Then the β's were fit by ridge regression to

$$E(Y|W) = \sum_{j=0}^{J} \beta_j E(X^j|W) + \sum_{j=1}^{K} \beta_{j+J} E\{(X - \xi_j)_+^J|W\}$$

MSE IN A SIMULATION OF FAN–TRUONG

Model	X	n	Str Spline	Simex	Decon Ker	Naive
$x^3(1-x)^3$	N	200	**2.1**	5.6	8.9	13.2
$x^3(1-x)^3$	S–N	200	**5.5**	6.6	10.9	14.6
$x^3(1-x)^3$	S–N	200	**5.6**	6.9	13.7	14.6
$\sin(4\pi x)$	U	500	**7.6**	14.5	20.3	31.2
$\{2-\sin(2\pi x)\}^{-1}$	U	500	3.8	**3.4**	6.2	7.6

bold = minimum MSE

- Distributions for X are
  - N = Normal
  - S–N = skew normal (badly modeled by mixture of normals)
  - U = Uniform (very badly modeled by mixture of normals)
- Methods are
  - Str Spline= Structural regression spline
  - Simex regression spline with quadratic extrapolant
  - Decon Ker = Deconvoluting kernel with best bandwidth over a range of bandwidths. (Fan & Truong)
  - Naive = Regression spline ignoring error
- Results with naive and Simex smoothing spline were slightly worse.

DISCUSSION

- Globally consistent nonparametric regression is ill–posed and essentially not practical.
- In many problems, the SIMEX approach gives nearly consistent estimates.
  - But not always!
- Structural estimation with penalized splines is flexible and convenient
  - Requires estimating the distribution of the unknown X's.
  - This can be done flexibly, by mixtures of normals or one's favorite method.