Parametric and non-parametric approaches to price and tax reform

Angus Deaton and Serena Ng

Research Program in Development Studies,
Princeton University

Boston College

This version: December 1997

(First preliminary version, March 1993)
(Revised January 1996, June 1997.)

Abstract

In many public policy problems, we need to estimate the way in which policy changes affect people's behavior. In the analysis of tax and subsidy reform, which is the topic of this paper, we need to know how tax-induced price changes affect the amounts that people buy of the taxed goods. We present various economic and statistical approaches to obtaining the estimates that are required. We consider the structural methods that are standard in economics, where the behavior and welfare of individual agents are simultaneously captured by the specification of utility functions whose parameters are to be estimated. We argue that these methods are less useful than alternatives that directly consider the derivatives of the regression function of average behavior. We consider both parametric and nonparametric estimators of these derivatives in the context of price reform for foods in Pakistan, focussing on the advantages and disadvantages of "average derivative estimation" (ADE) as proposed by Härdle and Stoker (1989) and Stoker (1991). Average derivative estimation is attractive in principle because it directly estimates the statistics that are required for policy analysis. In the practical case considered here, neither technique is a clear winner; each has strengths and weaknesses.

1. Introduction: policy change and behavioral responses

Much policy analysis is concerned with estimating people's behavioral response to a change in policy. In the case of tax and subsidy reform, which is the topic of this paper, we need to estimate how consumers respond to the changes in price that are brought about by the policy, for example how consumers in the US will change the amount they smoke in response to an increase in the price of tobacco, or how consumers in Pakistan will respond to a reduction in the subsidy on a staple such as wheat. A case can be made for changing (say raising) a price if the benefits of doing so exceed the cost. Estimating costs is relatively straightforward, at least for "small" price changes. The money value of the cost to a consumer of a unit increase in the price of a commodity is the quantity consumed; a 1 cent increase in the price of a pack of cigarettes costs 14 cents a week to someone with a two-pack-a-day habit. The benefits of the tax increase come from the social value of the additional revenue generated by the change. This revenue is usefully divided into two parts, the additional revenue that comes from the change in tax with behavior unchanged, and the change in revenue that comes from the change in behavior, at existing rates. It is the estimation of this last that poses the greatest challenge.

A simple economic framework can be used to make these ideas more precise. Suppose that the price of good i is p_i, and that this price contains a tax or subsidy element, so that we can write

$$p_i = v_i + t_i$$

where v_i is the price before tax, and t_i is the amount of the tax or subsidy. We suppose that t_i can be varied with v_i held constant, an assumption that must sometimes be modified, but which much simplifies a basic presentation. The costs of increasing t_i are conveniently written

$$\frac{\partial C}{\partial t_i} = \frac{1}{H} \sum_h \xi^h q_i^h.$$

C denotes a suitable measure of social costs which is averaged over the individual agents in the economy, labeled $h = 1,\ldots, H$. The derivative of the cost-of-living with respect to price *I* for individual *h* is the amount consumed, q_i^h, and the aggregate cost is obtained by weighting individuals by the amounts ξ^h, the purpose of which is to allow the possibility that policy makers weight different individuals differently, for example by giving greater weight to the poor, to local constituents, or to residents of a particular region. The benefits of the tax increase are represented by the change in revenue, *R*, again averaged over the population

$$\frac{\partial R}{\partial t_i} = \frac{1}{H} \sum_h q_i^h + \frac{1}{H} \sum_h \sum_k t_k \frac{\partial q_k}{\partial p_i}.$$

Note that the change in the price of *i* affects, not only the demand for good , but potentially also the demand for the other goods, $k = 1,\ldots, n$, and these effects must also be taken into account in the analysis.

The analysis of tax reform—Dixit (1975), Atkinson and Stiglitz (1980), and New-

bery and Stern (1987)—is concerned with the comparison of (2) and (3). The ratio of (2) to (3) can be interpreted as the social cost per unit of revenue raised through good i, so that if the ratio is low, the commodity is a candidate for a price rise, and if it is high, it is a candidate for a tax decrease or subsidy.

This paper is concerned with the estimation of these formulas. Equation (2) is straightforward; the weights ξ^h are matters for policy or political judgments, and not for estimation, and the quantities consumed by individual (or household) h can be obtained from household survey data which must also contain for each person the information necessary to evaluate the weights. Given that, the estimation of (2) is a routine exercise in survey sampling. The first term in (3) can be estimated from the same source, or from administrative tax records, so that it is only the last term that poses any difficulty.

The standard approach in economics—what we refer to here as the "structural" approach—is to specify a utility function for each agent h. This utility function tells us simultaneously how the welfare of h is affected by a price change, and how its demand is conditioned by price. Section 2 gives a brief summary of how the utility function can be specified, and the parameters estimated. It also identifies the main disadvantages of the approach, that the structural parameters—the parameters of preferences—are difficult to estimate without relying on arbitrary and largely un-

testable assumptions, and that, in any case, these parameters lead only indirectly to the quantities that need to be measured for policy. Section 3 considers an alternative strategy in which utility theory is largely ignored, where we start from a parametric specification of the regression function of average, instead of individual demands. We use this method to calculate results for the case study of this paper, which is the analysis of behavioral responses for tax and subsidy reform in Pakistan. Section 4 takes the analysis a step further, and uses average derivative estimators to provide direct estimates of the average behavioral responses. We discuss the advantages and disadvantages of these nonparametric estimates as compared with the parametric analysis in Section 3.

In a good deal of what follows, we simplify the presentation by focussing on one component of the second term on the right hand side of (3). This is the contribution of the own-price derivative to the change in revenue

$$t_i \frac{\partial \overline{q}_i}{\partial p_i} = \frac{1}{H} \sum_h t_i \frac{\partial q_i^h}{\partial p_i}.$$

Although it is sometimes important to consider cross-price derivatives—especially when there are closely related goods that are (or might be) taxed at different rates— (4) often accounts for most of the behavioral change. Note that (4) is typically nega- tive—demand curves slope downward—so that the larger absolutely is this term, the

smaller the revenue increase from the tax increase, and the higher the social cost of

raising additional revenue via good *i*. Other things being equal, goods for which (4)

is large and negative are goods where subsidies and taxes should be kept to a mini-

mum. Since the other things—(2) and the first term in (3)—are the same for the

different estimators considered in this paper, the policy consequences of different

estimation strategies are confined to the second term in (3), which will often be well

approximated by (4).

2. Structural approaches: utility-based demand systems

Although structural, preference-based analyses are used in a wide range of policy

exercises in economics, not only in public finance applications, but also in macro-

economics, we illustrate with tax reform in developing countries where much of the

literature, including that for Pakistan (see Ahmad and Stern 1991), assumes that

individual behavior conforms to some variant of the *linear expenditure system*

(LES), Stone (1954). Under this specification, household preferences are given by the

cost function (the minimum cost of attaining utility *u* at prices *p*)

$$x^h = c^h(u^h, p) = \sum_{k=1}^{n} \gamma_k^h \, p_k^h + u^h \prod_{k=1} ton \, p_k^{\beta_k}$$

where x^h is household total expenditure, u^h is utility, p_k is the price of good *k*, and

the γ_k^h and β_k ($\Sigma \beta_k = 1$) are parameters to be estimated, the former taken to be

household specific, the latter not. The demand functions associated with (5) are

$$p_i q_i^{\,h} = p_i \gamma_i^h + \beta_i (x^h \sum_{k=1}^{n} \gamma_k^h p_k)$$

When (5) is inverted to give utility as a function of prices and total expenditure, it

provides an explicit formula for the effects of tax induced price changes on individ-

ual welfare. Equation (6) is supplemented by an expression for γ_i^h, for example by

$$\gamma_i^h = \zeta_{i'} z^h + u_i^h$$

where z^h are vectors of observable household characteristics, and where the stochas-

tic term u_i^h captures unmodeled heterogeneity in demands. Given household level

data on demands, on prices and on total expenditures (or aggregate average data on

the same), together with a specification of u_i^h (or in the aggregate case, their aver-

age)—often multivariate normality—equations (6) and (7) can be taken to the data,

and the parameters estimated.

One immediate issue is the restrictiveness of the linear expenditure system. The

limited number of parameters in the model restricts the behavioral responses beyond

what is implied by utility theory. In data poor environments—as when using time-

series data in developing countries—this is an advantage. But in general, such a

specification largely begs the question with which we began; if individual house-

holds conform to the LES, then no matter what the values of the parameters, the cost-

benefit calculations will always approve a reduction in a tax rate that is higher than the average, and approve an increase in one that is below the average, see Atkinson (1977), Deaton (1987). In situations where the data are sufficient, the model can be replaced by one of several utility-based flexible functional forms. One such, which we refer to again below, is the almost ideal demand system of Deaton and Muell-bauer (1980), whose demand functions take the form:

$$s_i^h \; equiv \; \frac{p_i q_i^h}{x^h} = \alpha_i^h \; + \; \beta_i \; \ln\left(\frac{x^h}{P}\right) + \; \sum_{k=1}^{n} \theta_{ik} \; \ln p_k \, .$$

where P is an appropriately defined price index of all the prices. Formulations like (8) retain the utility basis of models like the LES, so that we retain the integration between welfare and demand systems, but they are much less restrictive and allow the data to affect the policy conclusions.

   A much more serious problem with both (6) and (8) is the difficulty in dealing with the fact that the quantities q_i^h (or expenditures $p_i q_i^h$) are censored at zero. Households do not report *negative* purchases, but in nearly all survey data, most households report zero purchases of at least some goods; indeed, all goods except basic staple foods have some zeroes. This phenomenon is inconsistent with either (6) and (7) or with (8) if u_i^h in (7) or α_i^h in (8) are continuously distributed. While the general theory of utility maximizing behavior allows for the possibility that some

goods are not purchased, its application does not lead to the straightforward func-

tional forms for demands as in (6) and (8). Instead, utility must be maximized with

the nonnegativity constraints explicitly imposed. In the simplest case, when $n = 2,$

and where one of the goods is always purchased, the demand for the other good can

often be handled as a censored linear regression model, as in Heckman (1974), but

with $n > 2,$ it is necessary to model the selection of goods—typically a polychoto-

mous choice problem—and to recognize that the functional form of the demand

function for each good will differ from one selection (or regime) to another. Even in

the simplest cases, structural modeling of this type has proved intractable, see Lee

and Pitt (1986). Even when $n = 2,$ a censored linear regression model is a good deal

more difficult to estimate credibly than is a linear regression model. The standard

maximum likelihood (Tobit) estimator is inconsistent under heteroskedasticity or

non-normality. While this can be repaired by the use of nonparametric Tobit

estimators, such as Powell's (1984) censored LAD estimator, implementations of the

structural approach beyond the empirically uninteresting two-good case, when

possible at all, requires the use of strong auxiliary assumptions about the distribution

of heterogeneity, assumptions that affect the estimates.

Note that, even if it were possible to estimate the structural parameters consistently

in the presence of censoring, it still remains to aggregate the individual derivatives up

to the average that is required for the policy analysis. To do so requires keeping track of which regime each consumer is in, averaging within each regime separately, and allowing for the effect of price changes on the regime selection.

The structural approach therefore presents us with two equally unpalatable alternatives. One is to estimate simple linear forms, such as (6) and (8), ignoring the censoring problem altogether, treating zero purchases in the same way as other purchases, or deleting them from the analysis. Either way, the theoretical structure—which is seen as the advantage of the approach—is compromised in the estimation, and the parameter estimates obtained will not be consistent for the preference parameters. Alternatively, we can follow the technically much more demanding route of modeling the regimes, and the behavior within each, so that we preserve the theoretical structure, but at the price of obtaining parameter estimates whose consistency is not robust to the changes in the auxiliary distributional assumptions that are required to obtain them.

3. Parametric estimation of average derivatives

Consider the estimation of the regression function

$$m_i(\ x\ ,\ p\ ,\ v\ ) = E(\ q_i / x,\ p,\ v)$$

where v is a vector of household variables on which we wish to condition, and which

are held constant (along with x) when taxes and prices are changed in computing the

derivatives in (3) and (4). If (9) can be estimated on household-level data, the deriva-

tives we need can be calculated from

$$\frac{\partial \bar{q}_i}{\partial p_j} = \int \frac{\partial m_i(x,p,v)}{\partial p_j} d F(x,v)$$

which can itself be estimated from the sample average of the calculated derivatives.

In the parametric work on Pakistan in Deaton (1997, Ch. 5), instead of working with

the regression functions of q_i and their derivatives, the authors work with the

regression functions of the *budget shares* $s_i = p_i q_i / x$, and, based on (8), adopt the

following two equation parametric form:

$$s_i^{ch} = \alpha_i^0 + \beta_i^0 \ln x^{ch} + \gamma_i^0 \cdot z^{ch} + \sum_{j=1} \theta_{ij} \ln p_j^c + (\; f_i^c + u_{0i}^{ch}\;)$$

$$\ln v_i^{ch} = \alpha_i^1 + \beta_i^1 \ln x^{ch} + \gamma_i^1 \cdot z^{ch} + \sum_{j=1} \ln p_j^c + u_{1i}^{ch}.$$

v_i is the "unit value" of good i, defined as the ratio of reported expenditure to re-

ported quantity (I spent 12 rupees, and bought 4 kilos, so that v is 3 rupees per kilo).

Although (11) looks similar to (8), it has a different interpretation; it is not a struc-

tural model of demand conditional on positive purchases, but an approximation to

the regression function. As such, and unlike (8), its parameters cannot be interpreted

as the parameters of preferences.

By equation (12), unit values are closely related to prices, which are not directly

observed, but because richer households buy higher qualities, unit values rise with x, and are possibly influenced by other household characteristics, z, as in (12). Where these effects are absent, $\ln v_i^{ch}$ and $\ln p_i^c$ are identical, the matrix Ψ in (11) is the identity matrix, and all other terms in (12) are zero. More generally, the matrix Ψ allows for "quality shading" in response to price; consumers may respond to higher price by buying, not only less, but also lower quality, so that when prices rise, unit values may rise by less.

The three levels of subscripts and superscripts refer to the good, i, the household, h, and the village (or sample cluster, or PSU) in which it lives, c. The unobservable prices, p_i^c are assumed to be constant within each cluster, f is an unobservable cluster fixed effect, and the u's are error terms that are allowed to be correlated across goods and between the share and unit value terms. Such correlation would be present, for example, if when asked to report expenditures and quantities, respondents calculated the former from quantity and prices, so that response errors in prices will be transmitted to reported expenditures.

The estimation of (11) and (12) is done in two stages. At the first, village effects are swept out by working with deviations from village means. This removes not only the fixed effects, but also the unobservable prices, so that the β and γ parameters can be consistently estimated. The correlations between the paired residuals in share and

unit value equations are calculated and interpreted as indicating the variance and

covariance of measurement error in the two equations. At the second stage, the

estimated β's and γ's are used to subtract out the demographic effects from (11) and

(12), and the "purged" budget shares and log unit values are averaged, village by

village. The between village regression of "purged" budget shares on "purged" log

unit values gives an estimate of price effects, which is corrected for the measurement

error calculated at the first step. As is apparent from (11) and (12), the matrices Θ

and Ψ are not separately identified, and in fact, the above procedure yields an

estimate only of $\Theta \ \Psi^1$, so that the estimation of Θ requires more information. This

is provided by a theory of quality shading developed in Deaton (1988, 1997, Ch. 5)

which uses a separability restriction on preferences that permits the two matrices to

be separately identified.

These are complex operations, and it is worth trying to assess which parts matter

in practice, and which might usefully be short-circuited. Start with the quality issue.

As first noticed by Prais and Houthakker (1955), these quality effects really do exist,

and there is nearly always a significant positive relationship between unit values and

total expenditure. However, the effects are not very large. The β-coefficients in (10)

are estimated to be 0.10 for both rice and wheat in rural Pakistan, they are a little

higher for dairy produce (0.14) and meat (0.15), but essentially zero for edible oils

and fats, and for sugar. Similar results were obtained for Indonesia, Côte d'Ivoire, and the state of Maharashtra in India, Deaton (1997). Numbers of this size make sense. It is hard to imagine a rich person paying much more than twice as much per kilo than a poor person for any broad aggregate of goods, and if rich people spend about six times as much as poor people—a useful rule of thumb —the elasticities will be of the size we have estimated. Given that the income elasticities of quality are so low, it is implausible that price elasticities of quality are high, which implies that the matrix Ψ is close to the identity matrix, and the final correction of the previous paragraph is not very important.

Even if quality effects can be ignored, measurement error remains a real hazard. In particular, evidence in Deaton (1988) suggested that regressing logarithms of quantities on logarithms of unit values is not a good idea. However, in (11) and (12), the transformation to budget shares and log unit values seems to remove the worst of the measurement error, at least in the sense that the correlations between the residuals of the first-stage within-village regressions are typically not very large, so that the correction for measurement error at the second stage also has little impact. There is no reason that this has to happen, although it is not implausible that while there are correlated measurement errors between reported quantities and reported unit values, there should be relatively little correlation in the reporting errors of expenditures and

unit values. Note also that the second stage regressions are run using village averages, so that the effects of measurement error, if not eliminated, are likely to be reduced.

The importance of the village fixed effects is harder to judge. While it is often the case that sweeping out the village means does not affect the estimates of β's and γ's—something that is true in Pakistan—the inclusion of village fixed effects seems like good practice. Villages often differ a great deal from one to another, and are frequently internally homogeneous, so that intravillage correlations are to be expected. Furthermore, the village specific factors, such as prices, are quite likely to be correlated with included characteristics such as income. Indeed, in earlier work with Côte d'Ivoire, there were several cases where inclusion or exclusion of village fixed effects had a marked effect on the first stage estimates.

Note that there is no explicit attempt here to deal with censoring. Zeros purchases are included in these regressions, as they must be for the derivatives to be averaged over all households, whether or not they consume each good. The estimation of equations (11) and (12) will give us the correct policy parameters, provided only that they correctly specify the regression functions; unlike the structural case, it is not required that the parameters correspond to any parameters of preferences, and in general they will not. Of course, there is no basis (other than convenience) for the assumption that

the regression functions take the linear form in (11) and (12).

Table 1 presents results from estimating (11) and (12) using data on seven important food groups from 9,119 rural households in Pakistan's 1984–85 Household Income and Expenditure Survey. We focus on the most important behavioral terms, the own-price responses, presented here in elasticity form. The wheat, rice, sugar, and edible oils groupings choose themselves because they are the groups whose prices were most distorted in Pakistan at the time of the survey, the other groups are chosen arbitrarily but conform to the usual decomposition of foods in Pakistani surveys. The second column shows the unrestricted own price elasticities, while the third column shows the same elasticities but with Slutsky symmetry imposed, both sets of estimates calculated as in Deaton (1997 pages 303-314). These are not very different, but we present them to show the ability of the structural approach to produce different estimates depending on how much of utility theory we wish to incorporate into the estimation procedures. The symmetric estimates use more of the theory, and in some cases (although not here) permit more precise estimates, or even the estimation of responses that cannot otherwise be measured. As incomes rise, the order of consumption of cereals in Pakistan is wheat and then rice, and this pattern is reflected in the price elasticities, with wheat inelastic, and rice—the relative luxury good— much more price elastic. (It is worth noting that similar estimates for the

state of Maharashtra in India, where wheat is the relative luxury and rice the more basic good, show a much higher price elasticity for wheat than for rice.) Conditional on the parametric specification, the standard errors are small, except for the two groups, oils & fats and sugar, whose prices are to some extent controlled, and which therefore display much less spatial price variation.

The price elasticities in the Table can be inserted into formulas (2) and (3) in order to calculate desirable directions of tax reform. In fact, the main recommendations are clear from the results and the formulas without detailed calculation. If we are interested in improving the lot of the poorest households, an increase in the price (reduction of the subsidy) on rice will have relatively low cost according to (2) because the poor do not consume rice, but will have relatively high benefit according to (3) because rice is highly price elastic, and the reduction in the subsidy will not only save resources directly, but will save additional resources as people reduce their consumption. Raising revenue through a reduction in the subsidy on rice is therefore estimated to have a high benefit to cost ratio. Because wheat is estimated to be a substitute for rice—the cross-effects were estimated but are not shown here—the beneficial effects of raising the price of rice would be moderated somewhat by the switch to wheat, which also carries a subsidy, but the effect is small relative to the equity and own-price effects.

4 Non-parametric approaches

The average derivative estimator to which we now turn requires, as its starting point,

regression functions in which average demands are conditioned on a list of variables,

including prices. In many policy applications, such a specification is the natural one

with which to begin. In the Pakistani example of this paper, the link with Section 2

can be made by starting from equations (11) and (12), and simplifying as follows.

Suppose that prices are equal to unit values, so that we assume away both measure-

ment error and quality issues. Suppose also that the fixed effects in (11) are uncor-

related with the observable conditioning variables, so that (11) implies

$$E(\ p_i q_i / x, p, z) = x(\ \alpha_i^0\ +\ \beta_i^0\ lnx\ + \gamma_i^0.z\ +\ \sum_{j=1}^{n}\theta_{ij}\ lnp_j\ )$$

The nonparametric approach dispenses with the functional form in (13), and esti-

mates its average derivative directly.

Average derivative estimation

To simplify the presentation, and for this section only, we denote the dependent vari-

able by *y*, which in the example will be the expenditure on each of the seven goods in

turn. We can then write, for each *i* in turn

$E(\ p_i q_i / z\ ) = E(\ y/z\ ) = m(\ z\ )$

and the quantity that we wish to measure is

$$b_j = E_z \left(\frac{\partial E(\ y|z\ )}{\partial z_j} \right) = E_z \left(\frac{\partial m(z)}{\partial z_j} \right).$$

The expression (15) is an *average derivative*, and to estimate it, we follow Härdle

and Stoker (1989) and Stoker (1991) and use *average derivative estimators*. This is

not the only way of proceeding—for example we could nonparametrically estimate

the regression functions themselves, using splines, kernels, or locally-weighted

estimators, and then calculate and average the derivatives—but they provide a

convenient, elegant, and direct method of estimation.

Average derivative estimation is sometimes thought of as a semi-parametric tech-

nique, perhaps because it is often used to estimate index regression models where the

conditional expectation of *y* is given by $\phi(\ x'\beta\ )$ for unknown function φ. In the

current context, however, we are not restricted to such models, and the method

produces estimates of average derivatives without having to restrict the functional

form. Even so, the rate of convergence of the estimators is much faster than is

typically the case for fully non-parametric treatments; like OLS, the estimates

converge at rate $n^{0.5}$, and not at $n^{0.2}$ as would be the case if we were estimating a

regression function or its derivatives. The more rapid convergence here is because we

are estimating only the *average* of the derivatives, so that although derivatives at any

given point will typically be estimated much more imprecisely, their average can be

relatively accurate.

   The theory of average derivative estimation is straightforwardly developed as

follows. Suppose that the joint density function of the conditioning variables z is

$f (z)$. Denote the "scores" by

$$w_j = - \partial \ln f (z)/\partial z_j$$

a quantity that, in principle, could be evaluated at each z^h in the sample, generating

an $H \ x \ K$ matrix, where K is the number of conditioning variables. Consider now

the unconditional expectation

$$E(w_j \ y)= \int \int w_j (z) \ y \ f (\ y, z \ ) d \ y \ d \ z = - \int \int f_j (z) \ y \ f (\ y/z \ ) dy \ dz$$

where the last expression comes from substituting (16) into (17), and then splitting

the joint density function into the product of a conditional and a marginal. If this last

term is integrated by parts, and if we assume that $f (z)$ is zero on the boundary,

then (17) becomes

$$E(w_j \ y)= \int \int \ f (\ z \ ) \ y \ f_j (\ y/z \ ) dy \ dz = E_z(\partial \ m(z)/\partial z_j \ ).$$

Hence, if we knew the w's in (16), and formed the $H \ x \ K$ matrix W, we could

calculate $\hat{b}_1 = H^{-1} W' y$

which, by (18) would converge to the vector of average derivatives (15). Note from

(18) that $E(w_j \ z_k)$ is the derivative with respect to z_j of the expectation of z_k

conditional on z, which is δ_{jk}, the Kronecker delta, so that the probability limit of

$H^{-1}W'Z$ is the identity matrix. In consequence, another consistent estimator of the average derivatives is provided by the "instrumental variable" estimator

$$\hat{b}_2 = (W'Z\ )^{-1}W'y$$

According to Stoker (1991), (20) is to be preferred to (19) because common biases in the denominator and numerator offset one another in a way that does not occur with (19).

   To make either of these estimators feasible, we require a method of estimating the scores (16). This first-stage estimation is based on a kernel estimate of the joint density $f\ (\ z\ )$ given by

$$\hat{f}\ (\ z\ ) = \frac{1}{\tau^k H}\sum_{h=1}^{\cdot\cdot} K\ \left[\ \frac{z - z^h}{\tau}\ \right],$$

where τ is a bandwidth and $K\ (\ .\ )$ is some suitable kernel function. In this paper, we use the quartic product kernel

$$K(\ z\ ) = \prod_{i=1}^{\wedge} \kappa(\ z_i\ ),\ \kappa(\ z) = \frac{15}{16}\ (\ 1 - |z|^2\ )^2\ \mathrm{I}\,(\,|z|\ \leq 1),$$

where $\mathrm{I}\,(.)$ is the indicator function that is 1 if the statement in brackets is true, and otherwise is zero. The bandwidth τ controls the degree of smoothing. Once a bandwidth is chosen—on which more below—the logarithmic derivatives of (21), which are the estimates of the scores (15), are calculated by differentiating (21) and using the data to calculate the resulting formula. This is straightforward but time consum-

ing; for each of the $H \; x \; K$, (in the Pakistani example below, 9119 by 9) elements of

the matrix W, the evaluation requires a complete pass through the sample of 9119

points. In practice, we follow the standard recommendations in Silverman (1986, p.

77–8), and first transform and scale the Z matrix so that it has a unit variance

covariance matrix, after which the calculated scores are transformed back to restore

the original dimensions. The bandwidth is chosen according to the recommendations

in Powell and Stoker (1996); in this case, we set τ to be unity for the transformed

data.

   Although it would be possible to estimate (20) using the estimate of the score

matrix as described, there will be problems where the estimated density is small,

since the evaluation of the logarithmic derivative requires division by the estimated

density. To avoid the problem, Powell, Stock and Stoker (1989) evaluate a density

weighted average derivative estimator, which corresponds to (15) weighted by

$f (z)$. But Powell, Stock and Stoker are only interested in estimating the average

derivatives up to scale, so that they can weight with impunity, whereas in the current

case, we need the average derivatives themselves. We therefore adopt an alternative

approach, "trimming" the data by deleting the 5 percent of the observations for which

the estimated density is smallest. If the cutoff on the density to achieve this 5 percent

is α, we can define the $Hx \, H$ diagonal matrix Ω by its typical element

$$\omega^{h'h} = \delta_{h'h} \ \mathrm{I} \, (\ \hat{\mathrm{f}} \, (\, \mathrm{z}^h) \ \mathrm{ge} \ \alpha).$$

Then the estimator that we actually use can be thought of as the weighted IV estimator

$$\hat{b}_3 = (\hat{W}' \, \Omega \, Z \ )^{-1} (\hat{W}' \, \Omega \, y \ ).$$

The estimates of the average derivatives are $\sqrt{H}$ consistent and asymptotically normal, with asymptotic standard errors given in (3.5) and (3.6) of Härdle and Stoker (1989) under the assumption that the observations are independent and identically distributed. Let the estimated "residuals" corresponding to household h for good i be given by:

$$\hat{r}_{hi} = \hat{w}_{hi} \ y_h \ \zeta_h \ + \ H^{-1} \tau^k \ \sum_{h'=1}^{H} [\ \tau^{-1} K'(\tfrac{z_h \, z_h'}{\tau}) \quad K(\tfrac{z_h \, z_h'}{\tau}) \hat{w}_{h'i} \,] \frac{y_{h'} \ \zeta_{h'}}{\hat{f}_{h'}}$$

Then the sample covariance matrix for the average derivative estimates is given by:

$$\hat{V}_0 = H^{-1} \sum_{h=1}^{H} \ \hat{r}_h \hat{r}_h \, \zeta_h \ \overline{r} \ \overline{r}'$$

Although this formula will be robust to heteroskedasticity in the residuals, it does not account for the possibly more serious bias to standard errors that comes from ignoring intracluster correlations, a particularly inappropriate omission in a context where the cluster structure is such an important part of the analysis. It is also known that the bias to the standard errors tends to be largest when the regressors vary little within the clusters, as is the case for the prices here, see Kloek (1981) and Pfefferman and

Smith (1985). The standard errors of the OLS estimates can readily be corrected

using a generalization of Huber-White procedures, and the results from our applica-

tion suggest that the problem is non-trivial, with the robust standard errors typically

more than twice the size of those reported. To adapt the standard errors of the ADEs

to allow for similar effects, we consider a modified sample variance-covariance

matrix, defined as:

$$\hat{V}_1 = H^1 \sum_{h=1}^{H} \sum_{h'=1}^{H} (\hat{r}_h - \bar{r})(\hat{r}_{h'} - \bar{r}) \, \zeta_{h'h}$$

where $\zeta_{hh'} = 1$ if household h and h' belong to the same cluster.

OLS as an alternative estimator

One immediate question is the relationship between ADE and OLS, with the com-

parison between the average derivatives in (15), as estimated by $\hat{b}_3$, and the coef-

ficients of the OLS regression of y on z. Although OLS estimates average slopes of a

sort, these are not generally consistent for the average derivatives in (15). In general,

if we place no restrictions on the shape of the regression functions then, as shown by

Stoker (1986), OLS will consistently estimate the average derivatives if, and only if,

the z vector is multinormally distributed. (Note that (20) is OLS if the scores are pro-

portional to the z's, which requires normality.) If the regression function is restricted

in some way, for example to be linear or quadratic, then the distributional assump-

tions on the *z*'s can be relaxed, but if we are unwilling to parametrize the model—which is the main point in the present case—OLS will not generally deliver consistent estimates of the quantities required for the policy analysis.

Figure 1, suggested to us by T. Stoker, shows an admittedly extreme case, but one that serves to illustrate that OLS and ADE estimation are different things. The true regression line has a sigmoid shape, and the distribution of *z* is bimodal, so that the data lie entirely within the two ellipses. The average derivative is the average slope over the top and bottom arms, while the OLS slope is as shown, and bears no relationship to the average derivative.

Some practical considerations

The non-parametric estimation procedure, like the parametric one, requires some way of dealing with those cases where there is no unit value recorded because the household in question did not buy the good. In the parametric estimation, price elasticities were estimated from the cross-cluster variation in unit values, effectively using average within-cluster unit values to represent each cluster. In the same spirit, we have filled in the missing values for each household with the average unit value for those households in the cluster who do make purchases. In the few cases where no one in the village records a purchase, we use the average for the province. We must

also check the condition for the consistency of ADE that the density of the explanatory variables is zero at the limits of the regression function. Our variables here are the logarithms of prices and the logarithm of per capita expenditure, which (in logs) appear to be roughly joint normally distributed, but with densities tending to zero at both high and low values. (Note also that the practical version of the estimator requires trimming of low density points.)

The calculations as described above are straightforward in principle; there are no implicit equations to be solved and no iterative techniques are required. The computation is burdensome only because there is a large number of observations H and a large number of conditioning variables, K, and because the number of evaluations is proportional to $K H^2$. The baseline calculations, using all observations, took over one day (over 200,000 seconds) to compute using a 8 processor (MIPS R4400) Unix workstation. This has prompted us to experiment with alternative methods to speed up the computations, all of which involve discarding some of the data. The first method is the simplest, and works by sampling a subset of the data randomly. The second and third methods make a more deliberate attempt to preserve the structure of the data. Both start by running a preliminary OLS regression of expenditure on the nine conditioning variables, and then ordering the observations by the size of the predicted values. This is done because we need a unidimensional quantity on which

to sort the data, but want to avoid selectivity bias induced by selecting on the dependent variable. Method two divides the sorted sample into bins (quintiles, deciles, for example), and then randomly select from each. Method three selects the fifth, tenth, or twentieth observation from the sorted data, starting from some observation other than the first.

Our experience suggests that the computation time is essentially determined by the effective sample size, and not by the method of sample reduction. Estimates from a reduced sample of 2,000 observations took one-third of the time required to obtain estimates from the full sample of 9,119 observations. The point estimates are quite robust to sample reduction. As might be expected, reducing the sample size raises the standard errors of the estimates, so that borderline significant estimates in the full sample tend to become insignificant in the smaller sample. However, estimates that are statistically significant in the full sample remain so in the smaller sample. There was no clear winner as to which of the three methods dominate in the sense of giving smaller standard errors. As these sample reduction methods have weak theoretical foundations, we are not particularly confident with these results, which we omit from this paper. Suffice it to mention that more sophisticated time saving methods based upon the idea of "discretization" and "convolution" are now being considered in the literature, Härdle and Linton (1994). We are, however, somewhat skeptical if such

methods are suited for high dimensional problems as here because the discretization procedure could itself be very time consuming .

We have nevertheless found a way of computing the average derivatives that results in a tenfold reduction in computing time. This comes from replacing the quartic kernel in (23) by the Gaussian kernel,

$$\kappa(z) = \frac{1}{\sqrt{2\pi}} \ e^{z^2/2}$$

Admittedly, we have not taken extra steps to optimize the execution speed of the programs, as this was not the intent of our paper. Indeed, we have been computing the average derivatives by "brute force." However, the average derivatives can be obtained in a mere 20,000 seconds (instead of 200,000) upon replacing the quartic kernel by the Gaussian kernel while leaving the rest of the program unchanged. Our cursory investigation into this difference suggests that the product Gaussian kernel involves one exponentiation and k addition operations. This kernel apparently entails drastically few computer instructions than the quartic kernel, which requires raising variables to higher power k times. Although not reported in published papers, this time-saving phenomenon has apparently been noted by others (T. Stoker in personal communication). In Table 2, we report the full sample estimates using the Gaussian kernel and the quartic kernel. When evaluated at the same bandwidth of 1.0, the Gaussian estimates are practically identical to the quartic estimates We report results

for the Gaussian kernel for $\tau=.75$ and $\tau=1.25$, which the readers can verify are sufficiently close to the quartic kernel with $\tau=1.0$. Clearly, for the application on hand, the estimates are robust to the choice of the kernel and the bandwidth.

Results and evaluation

Table 2 shows the bottom line, which is the comparison between the parametric esti-mates in Section 2, and the ADE of this section, with both sets of results presented as elasticities, in order to see more clearly the different policy implications. Wheat, rice, and sugar are all more price elastic according to the ADE calculations than under the original method, so that if these new results are correct, the subsidies on wheat and rice and the tax on sugar are more distortionary and less desirable than originally estimated.

   Table 3 shows the comparison between the OLS regression coefficients for each good, and the average derivative estimates for two kernels and different bandwidths. The most obvious feature of Table 3 is the close proximity of the OLS coefficients and the ADEs. Although there are a few differences, the two sets of estimates are never very far apart, and when the estimates are converted into elasticities in Table 3, for comparison with the parametric estimates of Section 2, the only difference of any importance is in the estimate for oils and fats, where the OLS elasticity is –1.31 and

the ADE elasticity is –1.77. The closeness of the OLS and average derivative esti-
mates is somewhat disappointing; the object of the paper was not to find the most
expensive possible way of calculating least squares! As we have seen, this is not a
general result. In the current case, the transformation of total expenditure to logs will
induce an approximate normality in that variable, but the prices are not obviously
normally distributed. Perhaps these results come from a combination of approximate
normality for the z's, and approximate linearity for the regression function in this
particular application.

The main distinction for the analysis of tax reform in Pakistan is not between OLS
and ADE, but between the original parametric method of Section 2, and the non-
parametric procedures, taking OLS and ADE together. We currently have no way of
deciding which set of estimates should be preferred. The original method deals with
issues—measurement error, quality effects, and village fixed effects— that we
ignored in the average derivative estimates, while the ADEs allow for more general
functional forms than permitted in the parametric model.

Consider first the *advantages* of the average derivative method. First, it offers a
direct measurement of the quantities that we need. The elasticities in the last column
of Table 3 do not come from evaluating the derivatives of some arbitrarily specified
functional form at some arbitrarily specified point. Rather they are estimates of the

second term in the denominator of the tax reform formula (3), and while the esti-
mates are interpretable as price elasticities, this is because it is convenient for the
purposes of comparison to think of them in such a way, not because we have chosen
to parametrize the elasticities. Second, the average derivative estimators do not
require that a functional form be specified for the demand functions. We do not have
to concern ourselves with utility theory, nor with the relationship between demand
functions and utility functions.

Third, we lose the multitude of problems associated with extensive and intensive
margins of consumption. It simply does not matter whether some consumers buy
some goods and not others and there is no need to treat zero demands any differently
from positive demands. The welfare and tax reform formulae do not require us to
treat zero purchases differently from positive purchases, but depend only on the
average consumed, and on the derivatives of these averages with respect to price.
Because ADEs estimate the average derivatives directly, and because their consis-
tency does not depend on whether the quantity being affected is zero or positive, or
has a zero or non-zero derivative, they can be applied directly to quantify the
formulas. In the structural approach, we would typically have functional forms for
the demand functions and their derivatives, conditional on positive demands, and the
calculation of the average derivative would come from integration over the fraction

of consumers making positive purchases. (For infinitesimal price changes, non-purchasers contribute nothing to the derivative, and demand is zero for those whose incomes and characteristics place them on the margin between purchasing or not purchasing at the current price level.) But there is currently no technique available within utility theory approach that enables us to calculate the set of prices, incomes, and household characteristics that correspond to positive purchases of each good. As a result, the integration is not feasible. On the other hand, the ADE technique renders it unnecessary. Of course, we should be careful about applying average derivative estimators to large price changes. To handle such cases, or to go beyond partial equilibrium analysis, there is no alternative to trying to estimate the demand functions themselves.

There are also a number of important *disadvantages*. The parametric model allows for village fixed effects in demands, thus recognizing that there are likely to be common but unobservable features of behavior in each village, and that these effects may well be correlated with village observables. Even when villages within an area have similar demand patterns, there are likely to be regional or provincial differences that are not simply attributable to differences in income and prices. Although we have not attempted to do so here, the ADE can readily be extended to deal with

regional effects, at least in theory. ADEs can be computed for each region and the results averaged using the appropriate regional sampling weights to get an estimate of the national average derivative that we require for the policy analysis. This method allows us to extend ADEs to allow for the presence of any categorical effect (dummy variable) among the explanatory variables, provided only that the number of observations within each category goes to infinity along with the overall sample size. The average effects of the dummy variables (ADE equivalents for finite samples) can also be obtained by calculating an average distance between regression functions for different values of the dummies, see the examples calculated in Deaton and Paxson (1998). However, it remains unclear in the current example whether ADEs can be extended to deal with fixed effects at the level of the village (PAU). Although ADEs can be computed for individual PIUS, and averaged, it is unclear whether such estimates will be consistent if—as is usually the case—PAU size remains fixed as the sample size grows. Estimators can certainly be constructed for special cases—such as additive fixed effects, or for index models with fixed effects, see Horowitz and Hardle (1994)—but these assumptions weaken the nonparametric argument for ADEs.

Secondly, the parametric estimation method has a procedure for treating the measurement error. While the effectiveness of the treatment is unclear, and there are

examples where the correction makes relatively little difference, the fact remains that within the parametric model there is a range of possibilities for dealing with measurement error, instrumental variables being the most obvious. The use of instrumental variables in semi-parametric applications is currently being developed, see Newey, Powell and Vella (1995) and Pinkse and Ng (1996), but the use of instrumental variables too deal with measurement error in linear models does not extend in any straightforward way to nonlinear models.

Thirdly, it is unclear how the quality correction procedures in the parametric model can be adapted to the nonparametric case. However as we have seen, these effects are typically small, and are unlikely to be a major source of difference between the two methodologies. If this were thought not to be the case, average derivative estimators could certainly be applied as far as the estimation of the elasticities of quality with respect to total expenditure. It is the next step, where the price elasticities are corrected, that has no obvious counterpart in the non-parametric case.

5. *Conclusion*

The ADEs solve some problems that are intractable in the parametric approach, and although their own problems seem to be addressable, there is a great deal more work to be done. There are still a number of difficulties that are common to both

approaches, and that are not resolved by the nonparametric techniques. In both

methods, it is necessary to specify a list of conditioning variables, and the results will

typically depend on the choice. There is therefore no possibility of a fully non-

parametric treatment, so that the estimation of the behavioral responses in (3) is still

on a very different footing from the estimation of the means in (2). Second, the

problem of missing unit values is still largely unresolved. In the implementation of

the ADEs we imputed prices to clusters on the basis of geographical information, a

procedure that is obviously sensible and that is supported by the good fit that is

obtained when observed unit values are regressed on cluster dummies. However, the

parametric approach, by sweeping out cluster fixed effects, requires prices only at the

cluster level, so that less imputation is required. Furthermore, it seems that different

imputations schemes give different results, so that, for example, the OLS estimates in

Table 2 are changed non-trivially if regressions are run on cluster means rather than

on the individual data. It could also be argued that the need for a parametric

imputation technique for missing values is a good deal less comfortable in a

nonparametric setting than in a parametric model. It is also possible that "automatic"

imputation techniques could be developed in the nonparametric context. Although

there are still very real difficulties, we feel that the use of average derivative

estimators and semi-parametric methods to analyzing tax reforms are sufficiently

promising to reward that work.

List of works cited:—

Ahmad, S. Ehtisham, and Nicholas H. Stern, 1991, *The theory and practice of tax reform in developing countries*, Cambridge. Cambridge University Press.

Atkinson, Anthony B., 1977, "Optimal taxation and the direct versus indirect tax controversy," *Canadian Journal of Economics*, **10**, 590–606.

Atkinson, Anthony B., and Joseph E. Stiglitz, 1980. *Lectures on Public Economics*. New York. McGraw-Hill

Deaton, Angus S., 1987, "Econometric issues for tax design in developing countries," in Newbery and Stern, 92–113.

—— 1988, "Quality, quantity, and spatial variation of price," *American Economic Review*, **78**, 418–30.

—— 1989, "Household survey data and pricing policies in developing countries," *World Bank Economic Review*, **3**, 183–210.

—— 1997, *The analysis of household surveys: a microeconometric approach to development policy*, Baltimore, Md. Johns Hopkins U. Press for the World Bank. (forthcoming.)

—— and Christina H. Paxson, 1998, "Economies of scale, household size, and the demand for food", *Journal of Political Economy,* (forthcoming).

Dixit, Avinash K., 1976, "Welfare effects of tax and price changes," *Journal of Public Economics*, **4**, 103–23.

Härdle, Wolfgang, and Thomas M. Stoker, 1989, "Investigating smooth multiple regression by the method of average derivatives," *Journal of the American Statistical Association*, **84**, 986–95.

——— and Oliver Linton, 1994, "Applied Nonparametric Methods," in Engel, Robert F. and Daniel F. McFadden, eds., *Handbook of Econometrics*, Volume IV, Amsterdam. North Holland.

Heckman, James J., 1974, "Shadow prices, market wages, and labor supply," *Econometrica*, **42**, 679–94.

Horowitz, Joel L. and Wolfgang Hurdle, 1994, "Direct semiparametric Estimation of single-index models with discrete covariates", mimeo, University of Iowa.

Kloek, Teun, 1981, "OLS estimation in a model where a microvariable is explained by aggregates and contemporaneous disturbances are equicorrelated," *Econometrica*, **49**, 205–7.

Lee, Lung-Fei and Mark M. Pitt, 1986, "Microeconometric demand systems with binding non-negativity constraints: the dual approach," *Econometrica*, **54**, 1237–42.

Newbery, David M.G., and Nicholas H. Stern, 1987, *The theory of taxation for*

developing countries, Oxford. Oxford University Press for the World Bank.

Newey, W. K., J. L. Powell, and F. Vella, 1995, "Non parametric estimation of triangular simultaneous equations models", Rice University Discussion paper 95-13.

Pfefferman, D., and T. M. F. Smith, 1985, "Regression models for grouped populations in cross-section surveys," *International Statistical Review*, **53**, 37–59.

Pinkse, J. and S. Ng, 1996, "Non parametric two step estimation of unknown regression   functions when the regressors and the regression error are not independent", C.R.D.E. Discussion paper 1596, University of Montreal.

Powell, James L., 1984, "Least absolute deviations estimation for the censored regression model," *Journal of Econometrics*, **25**, 303–25.

—— James H. Stock, and Thomas M. Stoker, 1989, "Semiparametric estimation of index coefficients," *Econometrica*, **57**, 1403–30.

—— and Thomas M. Stoker, (1996) "Optimal bandwidth choice for density weighted averages," *Journal of Econometrics*, 75, 291-316.

.Prais, Sigbert J., and Hendrik S. Houthakker, 1955, *The analysis of family budgets*, Cambridge. Cambridge University Press. (2nd ed. 1972)

Silverman, Bernard W., 1986, *Density estimation for statistics and data analysis*, London and New York. Chapman and Hall.

Stoker, Thomas M., 1986, "Consistent estimation of scaled coefficients," *Econometrica,* **54**, 1461–81.

Stoker, Thomas M., 1991, *Lectures on semi-parametric econometrics*, Louvain-la-Neuve, Belgium. CORE Foundation.

Stone, J. Richard N., 1954, "Linear expenditure systems and demand analysis: an application to the pattern of British demand," *Economic Journal,* **64**, 511–27.

Timmer, C. Peter, 1981, "Is there `curvature' in the Slutsky matrix?" *Review of Economics and Statistics,* **63**, 395–402.

Timmer, C. Peter and Harold Alderman, 1979, "Estimating consumption parameters for food-policy analysis," *American Journal of Agricultural Economics*, **61**, 982–4.

Table 1
Own-price elasticities for seven goods
Rural Pakistan 1984–85

	budget share	unrestricted	symmetric
wheat	0.1276	–0.61 (0.10)	-0.63 (0.10)
rice	0.0267	–2.16 (0.25)	-2.04 (0.22)
dairy produce	0.1269	–0.89 (0.04)	-0.90 (0.04)
meat	0.0366	–0.57 (0.18)	-0.54 (0.18)
oils & fats	0.0414	–0.80 (0.18)	-0.81 (0.18)
sugar	0.0293	–0.11 (0.53)	0.09 (0.53)
other food	0.1219	–0.51 (0.10)	-0.50 (0.09)

Notes: From Deaton (1997). Unrestricted estimates of price elasticities are given in Column 2 and elasticities with Slutsky symmetry imposed are reported in column 3.

Table 2
Own-price elasticities from OLS and ADE estimates
Rural Pakistan 1984–85

	Parametric	OLS	ADE
wheat	-0.61 (0.10)	–0.93 (0.02)	–0.91 (0.03)
rice	–2.16 (0.25)	–2.31 (0.06)	–2.31 (0.07)
dairy produce	–0.89 (0.04)	–1.05 (0.01)	–1.09 (0.02)
meat	–0.57 (0.18)	–0.20 (0.04)	–0.20 (0.06)
oils & fats	–0.80 (0.18)	–1.31 (0.14)	–1.77 (0.17)
sugar	–0.11 (0.53)	–0.30 (0.22)	–0.26 (0.17)
other food	–0.51 (0.10)	–0.51 (0.03)	–0.61 (0.04)

Notes: The first column is the second column of Table 1. The second and third columns are calculated from the first and second columns of Table 2.

Table 3
OLS and Average Derivative estimates of expenditures with respect to log prices
Rural Pakistan 1984–85

	OLS	ADE Quartic kernel $\tau=1.0$	ADE Gaussian kernel $\tau=1.25$	ADE Gaussian kernel $\tau=0.75$
wheat	11.71 (3.9)	14.33 (5.4) [8.6]	14.39 (5.4) [8.6]	14.16 (5.4) [8.6]
rice	–46.12 (1.9)	–46.08 (2.6) [5.9]	–46.07 (2.6) [5.9]	–46.08 (2.6) [5.9]
dairy produce	–8.30 (2.4)	–14.97 (3.8) [7.5]	–14.26 (3.9) [7.7]	–14.96 (3.8) [7.5]
meat	38.30 (2.1)	38.72 (3.0) [5.0]	38.79 (3.0) [5.0]	38.72 (3.0) [5.0]
oils & fats	–16.66 (7.4)	–42.06 (9.4)[19.6]	–42.10 (9.4)[19.6]	–42.06 (9.4)[19.6]
sugar	26.76 (8.7)	28.44 (6.7)[13.4]	28.41 (6.7)[13.4]	28.44 (6.7)[13.4]
other food	79.08 (4.4)	62.77 (6.5)[11.5]	62.77 (6.5)[11.5]	62.77 (6.5)[11.5]

Notes: OLS is ordinary least squares using all 9,119 observations. ADE is average derivative estimation, τ is the bandwidth. Heteroskedastic consistent standard errors are in rounded brackets; standard errors corrected for cluster effects are in square brackets.