

Consistent Policy Analysis when Some Heterogeneity is Unobserved

AMIL PETRIN
*Graduate School of Business
University of Chicago
and NBER*
December 1, 2003

PRELIMINARY/INCOMPLETE

Abstract

The answers to policy questions often depend on knowledge of the entire distribution of observed and unobserved factors entering a model. The reason is straightforward: the policy functional is rarely linear in its arguments, so unobserved heterogeneity does not average out. The preeminent role in economics of transforming skewed dependent variables using the log function suggests that applications may suffer from this bias. This paper describes this bias and provides methods to alleviate it. A major theme is that jointly determined variables contain information on common unobserved factors. The empirical application exploits the simultaneous determination of cable television adoption and the amount that it is watched to uncover otherwise unobserved heterogeneity. For the linear, log-linear, log-log, and logit demand models, the distribution of welfare changes induced by a large price increase differs substantially when the utilization information is added. The median welfare change decreases by on average almost 20%. For some infra-marginal markets, the welfare change is misstated by between 50% and 80%.

*Financial support from the National Bureau of Economic Research is greatly appreciated.

Correspondence to: 1101 E 58th St. Chicago, IL 60613.
email: amil.petrin@gsb.uchicago.edu.

1 Introduction

Estimates of demand, supply, and production are the fundamental input to applied welfare analysis. The relevant units for policy analysis are given by the levels of the dependent variables in these estimated equations, whether it be quantity, wage, or output, for example. When the statistical model is not linear in the levels of the dependent variable, observed and unobserved factors affect the distribution of outcomes in a non-linear manner. Perhaps the most common example in economics is the log transformation, which is a favorite for skewed dependent variables,¹ and is almost universally applied in policy analysis that uses regressions of (log) quantity on price and other controls, (log) wage on factors like education and ability, and (log) output on inputs for production functions.

In the case of such non-linearity, the full statistical model must be used to evaluate the distribution of changes induced by the policy. Consistent counterfactual forecasts of the policy's effects are constructed by first (re)transforming the estimated equation so that predictions from it are in the original units of the dependent variable (the relevant units for the analysis). Then, the policy effect is calculated for each agent, as different types will experience different changes. This yields the distribution of changes, which is then usually aggregated into one or several averages that is used determine the fate of the proposed policy (or its effect after it has been implemented). Following this approach is of particular importance when the environmental change is not "infinitesimal," and when the weights in the aggregation are focused heavily on subgroups that are infra-marginal.²

On its face, this calculation appears deceptively simple. The difficulty in implementation becomes apparent when one recognizes that the re-transformed

¹In these cases it can confer certain statistical benefits, like linearity and/or homoskedasticity, which makes unbiased and consistent estimation of parameter estimates and their standard errors straightforward.

²Many questions revolve around more substantial changes in the economic environment, like the surplus generated by the introduction of a new product or the increase in earnings arising from reducing class sizes by 30%. Many questions are also interested in non-representative individuals or groups.

model is not linear in the relevant unobserved factors. As Hausman and Newey (1995) note, this non-linearity means that “one can ignore the residual if it is all measurement error but not if it contains individual heterogeneity.” Put another way, all relevant factors must be accounted for (and all measurement error must be ignored) when the policy analysis takes place because the relevant unobserved factors do not average out. Applications where this bias arises are common in economics, including estimation of consumer/producer surplus, the effect on earnings of schooling inputs or military service, and questions regarding GDP and productivity growth.

This paper explains the theory behind the bias. When the policy functional is not linear in its arguments, all unobserved heterogeneity must be accounted for when estimating the effect of a policy change. The problem for the econometrician is that he or she cannot distinguish in the estimated error the part attributable to unobserved heterogeneity.

This paper provides methods to distinguish in the estimated error the part attributable to unobserved heterogeneity, so the full distribution of the relevant observed and unobserved outcomes is available (and can be used in the policy analysis). These methods recover unobserved factors using observed outcomes. The theme of the methods is that different outcomes determined in part by the same underlying unobserved factors contain information that allow one to separate out the unobserved factor from the residual. Using insights from Madansky (1964), estimates of the unobserved factors contribution to any of the related equations can be constructed.³ A number of different cases common to empirical work are analyzed.

The application exploits the fact that, for many goods, a consumer’s valuation is reflected in two decisions: whether the good is adopted and how much it is used. These two decisions are made simultaneously and typically reflect some underlying factors that are similar and some that may differ. The application’s design applies to durable good expenditures, which account for over

³There is a growing literature that uses control functions to obtain consistent estimates of model parameters associated with observed factors (see Heckman and Robb (1986), Pudney (1981), Pudney (1982), and Heckman and Scheinkman (1987), for example. The emphasis here is on estimating the unobserved factors.

30% of total consumer expenditures.⁴ Examples include the purchase of a car and the mileage it is driven, and the adoption of household appliances (like air conditioners or heaters) and the use of electricity or gas to power them. It also applies to any good that has a two part tariff pricing structure, like land-line and cell phone use.

The empirical application in this paper uses data from the telecommunications industry. Many policy questions arise in this industry regarding the effects of mergers (AOL/Time-Warner in 2000, ATT/Comcast in 2001, and EchoStar-DirecTV in 2001), new product introductions (Satellite dish, DSL and cable modem internet access), regulation of cable franchise monopolies (price deregulation, promotion of entry), programming, and other related issues. The application here uses variability in the amount of television watched to improve demand estimates for the adoption of cable television. Two individuals that look the same “demographically” may differ quite substantially in their unobserved tastes for television, and their amount of television viewing in part reflects this otherwise unobserved difference.

To show the relevance of the methods for an applied case, the application uses market-level data, the most commonly available type of data, to evaluate the change in welfare from a large price increase in the expanded basic cable price. In particular, market-level averages of television viewing are shown to have, conditional on other market-level observables, explanatory power in the market-level adoption equation for cable television. For the linear, log-linear, log-log, and logit demand models, the distribution of welfare changes induced by a large price increase differs substantially when the utilization information is added. The median welfare change decreases by on average almost 20%. For some infra-marginal markets, the welfare change is misstated by between 50% and 80%.

⁴Research on demand estimation has recognized the simultaneity of a durable’s adoption and utilization since the introduction of discrete/continuous models by Hausman (1979), Dubin and McFadden (1984), Hanemann (1984), and Mannering and Winston (1985). They describe related choices made by consumers that have both discrete (the adoption) and continuous (the use) aspects to them. In these cases the discrete and the continuous choice are determined simultaneously by some underlying factors that are similar and some that may differ.

Sections 2 explains the bias problem when heterogeneity is not accounted for and section 3 provides examples of this that span a wide range of empirical applications. Section 4 explores the econometrics. Section 5 contains the application. Section 6 concludes.

2 The Bias Problem

We describe how the bias arises. The relevant factors for any observation are given by $(Y_i, Z_i, \omega_i, \varepsilon_i)$, where Y_i is a vector of observed endogenous variables, Z_i is a vector of observed controls, ω_i is a (possible vector) of unobserved “systematic” factors, and ε_i is an unobserved component of the error arising because of measurement problems (say). One endogenous variable is of particular interest for the policy question. Define it as Y_{i1} . Its observed value is a function of the other endogenous variables Y_{i-1} , exogenous factors, and components of the error ω_i and ε_i , and is given as

$$Y_{i1} = f(Y_{i-1}, Z_i, \omega_i, \varepsilon_i; \theta, \lambda),$$

where Y_{i1} written here is measured in levels. Finally, θ and λ are (possibly vectors of) parameters, with λ defined as the parameters on the unobserved factors ω_i and θ used for endogenous and exogenous variables.

The answer to the policy question is based on evaluating $f(\cdot)$ at (possibly many) different values of its arguments. In these calculations, ω_i is defined as the part of the error that needs to be incorporated into the analysis. It arises because certain relevant factors like “ability” or “taste” are not observed. The additional component of the error, ε_i , is defined as the part of Y_{i1} attributable to factors the researcher would like to exclude from the analysis (arising from measurement problems, for example). With ε_i excluded from the calculation (i.e. evaluated at its population average of zero),

$$f(Y_{i-1}, Z_i, \omega_i, 0; \theta, \lambda)$$

gives the predicted value of Y_{i1} , in levels, that plays the relevant role in evaluating the welfare/policy implications of an environmental change.

Let $W(\cdot)$ denote the relevant operation on $f(\cdot)$ that yields the policy estimate.⁵ Then, for any agent the relevant computation is given by

$$W(f(Y_{i-1}, Z_i, \omega_i, 0; \theta, \lambda)).$$

For example, when a measurement of consumer surplus is desired, $f(\cdot)$ represents demand (in levels) and the surplus operation is given by integrating the demand function over price, holding Z_i, ω_i constant, and excluding ε_i from the list of demand shifters.⁶ The result for the aggregate is based on the integral over the population, or

$$\int_{Z, \omega} W(f(Y_{-1}, Z, \omega, 0; \theta, \lambda)) dP(Z, \omega) \quad (1)$$

where $dP(Z, \omega)$ gives the distribution in the population of observed and unobserved factors (or the weights applied to them in a policy aggregate), and ε is held constant at its mean value of zero during the integration.

The bias in policy estimates arises because the distribution of ω is not typically known, that is, it is hard to distinguish in the estimated residual the relevant heterogeneity from the measurement-type error ε . The standard approaches take one of two “extreme” viewpoints. One method ignores the entire error, setting both parts of the residual to their population averages of zero when the policy analysis is undertaken. This policy estimate is given by

$$\int_Z W(f(Y_{-1}, 0, 0; \theta, \lambda)) dP(Z) \quad (2)$$

where ω is (like ε) evaluated at its mean of zero, instead of being integrated over as in (1). The second approach is to add back the entire error, so the aggregate estimate is given by

$$\int_Z W(f(Y_{-1}, \omega, \varepsilon; \theta, \lambda)) dP(Z) \quad (3)$$

⁵If it is also not linear in $f(\cdot)$, this only exacerbates the bias problem.

⁶How one treats the prices of substitutes, which also typically enter as endogenous variables, will depend upon the question at hand. In some cases, one will want to allow the endogenous variables to vary in the functional, while in other cases one will want to hold them constant.

The non-linear manner in which ω enters $f(\cdot)$ (and thus $W(\cdot)$) means neither (2) nor (3) generally equals (1). Put another way, the mistakes do not average out. The difference between (1) and either (2) or (3) is the bias. This is the problem, and often there is nothing about the economics of the situation to suggest this bias is small. In fact, given the low explanatory power of some regressions, the econometrics of the situation suggests the bias could be very large.

The most popular transformation – the natural log – provides an illustrative example. The transformed the dependent variable is written as linear in regressors, parameters, and error, or:

$$\log(Y_i) = Z_i\theta + \lambda\omega_i + \varepsilon_i,$$

so linear techniques can be used in the estimation of the parameters θ . To make the point simply, assume $\theta = 0$ and ω is distributed normally with mean zero and variance σ^2 . The correct approach (equation (1)) integrates e^ω over the distribution of ω , or $\int_\omega e^\omega dP(\omega)$. The normality assumption means this integral equals $e^{\sigma^2/2}$. The approach that ignores all of the heterogeneity (equation (2)). applies the exponential function to the random variable ω evaluated at its mean value, or e^0 , and then integrates across the population (since $\int_\omega dP(\omega) = 1$, $\int_\omega 1dP(\omega) = 1$). In this case, the only time (2) is correct is when there is no taste heterogeneity, so ω is a degenerate random variable (i.e. $\sigma_\omega^2 = 0$). As σ_ω^2 increases, (with the mean held constant at zero), the bias becomes worse. This example suggests using (2) can result in large bias when little of the variance in the dependent variable is explained in the regression because relevant factors are unobserved. A similar problematic story can be told for (3).

3 Some Common Cases

The following is a partial list – meant to be illustrative – of the wide range of cases when this bias problem arises:

(i) *Consumer (or producer) surplus* – When the policy question requires an estimate of consumer surplus, demand estimation that recognizes hetero-

geneity among agents is usually the important input. A simple example of the bias arising from ignoring heterogeneity is provided figures 1a and 1b. Two types of agents face an increase in price for a good they consume. The high demand type (A) and the low demand type (B) differ only in their intercept; their slope's are assumed to be identical in $\log(q)$ - p space (figure 1a). The econometrician, ignoring the heterogeneity because it is unobserved or because it does not affect the consistency of the demand estimates from the $\log(q)$ - p equation - estimates the "representative" consumer's demand curve as line C. The aggregate surplus change induced by a price increase from p_0 to p_1 is computed using the levels equations in figure 1b (which are the transformed log-linear demand curves from figure 1a and given by *). Surplus equals the sum of areas under A^* and B^* and is equal to $D+E+2F$. The representative consumer model uses twice the area under C^* to estimate welfare, which equals $2E+2F$. The error in the surplus estimate is $D-E$. This error worsens as the non-linearity becomes more severe, and it does not converge to zero as the sample size increases.

As Hausman and Newey (1995) succinctly put it, for a large class of demand estimates (e.g. AIDS models and log-log models), ignoring heterogeneity is "consistent with current practice in applied econometrics, and is difficult to improve without more information about the residual." For discrete choice models, the opposite assumption typically holds, where the *entire* error is assumed to be utility and thus enters any policy calculation.

(ii) *The impact of schooling inputs on earnings* – The impact of schooling inputs on future earnings is at the heart of education policy.⁷ The policy analysis is often based on an educational production function which relates the log of wage to the inputs of schooling and other controls. From a policy planner's perspective, a necessary condition for policy approval is that increases in future (depreciated) earnings from additional schooling inputs exceed the costs of providing the additional inputs. The role of the estimated wage function

⁷These include investigations of the effects of class size, teacher quality, available resources, and the like. See, for example, Mincer (1974), Griliches (1977), the 1998 issue of the Review of Economics and Statistics, and the ongoing debate on class size between (among many others) Hanushek and Krueger (see Krueger, Hanushek, and Rice (2002)).

is to forecast the counterfactual wage, and since wage functions are almost always estimated in log-levels, these forecasts require the full model. In particular, when the changes in schooling inputs are large – those often considered by policy makers – all observed and unobserved factors (e.g. ability) must be accounted for during the re-transformation to wage levels because mistakes in the forecasted wage levels do not average out in the aggregation. In the same way, these biases arise when measuring the the effects of military service on earnings.

(iii) Productivity growth – Understanding the effects of policies on firm input choices and productivity growth is fundamental to comprehending their effects on GDP growth and its prospects. These questions require estimates of firm production functions, which are then used to forecast the effects of the policy changes. The potential for bias in these estimated effects is large for two reasons. First, production functions are almost always estimated in log-levels. Second, the recent focus on firm-level data has brought to light substantial heterogeneity in output at the firm-level even after conditioning on skilled and unskilled labor, capital, materials, and energy inputs. Since the unobserved heterogeneity enters the levels equation for output non-linearly, it must be conditioned on when counterfactual predictions for output and revenue are computed.

4 Alleviating The Bias

This section outlines methods for recovering unobserved factors. The idea is to use the errors from other related outcome variables that also have the same unobserved factors as explanatory variables. The focus is on the cases when the relevant equations can (after transformation, say) be represented by a linear system, although the theme of ideas discussed extends directly to non-linear cases.

The goal is to construct an alternative estimator for ω_i 's contribution to the equation of interest. Call the estimate $\hat{\omega}_i$. Letting $dP(Z, \hat{\omega})$ denote the joint distribution of observed factors and estimated ω_i 's, our proposed policy

calculation is given as

$$\int_{Z, \hat{\omega}} W(f(Y_{-1}, Z, \hat{\omega}, 0; \theta, \lambda)) dP(Z, \hat{\omega}). \quad (4)$$

We define an estimator for $dP(Z, \hat{\omega})$ and show it is consistent as the number of observations on $\hat{\omega}_i$ increases to infinity, that is, as the number of equations per agent increases to infinity.

This raises an incidental parameters problem posed by Neyman and Scott (1948). In applied cases the asymptotics will almost always be from the number of agents increasing with the number of equations per agent fixed. This case is similar to the well-known fixed effects case with panel data, where only an unbiased estimate of the fixed effect can be constructed when the time-series is small.⁸ The appeal is that the calculation given by (4) provides an estimator with potentially much less bias than that provided for by the two extreme approaches given by (2) and (3), where essentially $\hat{\omega}_i = 0 \forall i$ or $\hat{\omega}_i$ equals the full residual.

The discussion in this paper will assume that identification holds for the equations of interest and directly address methods for estimation of unobserved factors.⁹ For some cases this will not require further assumptions than those already used in the applied literature to estimate the principle equation of interest.¹⁰ However, further assumptions are necessary to estimate the additional equations used to identify ω_i 's contribution to the equation of interest. The benefit is that one is not left in the position of relying exclusively on (2) or (3), both of which are inconsistent, and, even taken together, can provide no definitive answer to how different the underlying distribution of unobserved heterogeneity is from the assumed distributions in (2) or (3).

⁸In fact, the fixed effects case is a special case of the approach we describe that arises when the same outcome variable is observed repeatedly and the contribution of the unobserved factor is constant over time.

⁹There is a voluminous literature on identification in simultaneous equations systems. See, for example, Chapters 4 and 7 in the Handbook of Econometrics, Volume One, and the literature cited therein.

¹⁰In some cases one will want to consider a structural equation where previously a reduced form equation may have been estimated, as the application illustrates.

4.1 Two Equations/One Unobserved Factor

Consider the two equations Y_{i1} and Y_{i2} :

$$Y_{i1} = Z_{i1}\theta_1 + \theta_{1,2}Y_{i2} + \epsilon_{i1} \quad (5)$$

$$Y_{i2} = Z_{i2}\theta_2 + \epsilon_{i2}, \quad (6)$$

where Z_i 's denote the observed exogenous factors and the θ 's their associated parameters, and the errors decompose as

$$\epsilon_{i1} = \lambda_1\omega_i + \varepsilon_{i1} \quad (7)$$

$$\epsilon_{i2} = \lambda_2\omega_i + \varepsilon_{i2}, \quad (8)$$

where ω_i is the common unobserved factor (with coefficients λ_1 and λ_2) reflecting “shifters” that must be accounted for in the analysis, and ε_i 's reflect mean zero independent errors that do not enter the policy analysis. For example, some characteristic's based demand approaches write demand (Y_{i1}) as a function of demographics and observed characteristics Z_{i1} , price Y_{i2} , and an unobserved product characteristic (ω_i), where i indexes markets (say). The equilibrium pricing function is often written as just a function of the exogenous variables (including the unobserved characteristic).¹¹

Consistent estimation of the first equation requires at least one observed factor from Z_{i2} is excluded from Z_{i1} , and that the unobserved factor is uncorrelated with Z_{i1} . For the example, this requires an instrument like a cost shifter, and that the unobserved characteristic be uncorrelated with observed characteristics, a standard (although perhaps objectionable) assumption made in the literature. For any welfare computation involving this demand curve, the true distribution of demands requires conditions on $\lambda_1\omega_i$, the unobserved product characteristic that varies across markets and affects willingness to pay.

More generally, one can consider the two equation system of simultaneously determined variables Y_{i1} and Y_{i2} :

$$Y_{i1} = Z_{i1}\theta_1 + \theta_{1,2}Y_{i2} + \epsilon_{i1} \quad (9)$$

$$Y_{i2} = Z_{i2}\theta_2 + \theta_{2,1}Y_{i1} + \epsilon_{i2}, \quad (10)$$

¹¹This will typically be a function of all the product characteristics in the market, the sellers, etc...

with the same error structure as above. For example, Griliches (1977) (summarizing the literature) writes schooling's (Y_{i2}) affect on (log) earnings (Y_{i1}) in this form. The well-known simultaneity problem arises because ability, represented by ω_i , is typically unobserved and affects both schooling and earnings. It is not assumed, however, to be correlated with the other determinants in Y_{i1} . Griliches discusses the possibility that the (expected) wage may in part determine the amount of schooling. Consistent estimation in this case requires that at least one observed factor from Z_{i1} is excluded from Z_{i2} (and vice versa for the second equation). For policy questions, the change in the distribution wages induced by a policy change that increases schooling (say) will in part determined by the effect of an individual's ability on wage (via $\lambda_1\omega_i$).

4.2 Recovering $\lambda_1\omega_i$

To illustrate the idea for recovering $\lambda_1\omega_i$, we begin with a simple (and unrealistic) case, assuming that all of the error entering the second equation is “systematic” (i.e. $\varepsilon_2 = 0$). In this case This means that the ratio $\frac{\lambda_1}{\lambda_2}$ is consistently estimated from the ratio of covariance between ϵ_{i1} and ϵ_{i2} to the covariance of ϵ_{i2} with itself (i.e. its variance), or

$$plim \frac{\sum_{i=1}^N \epsilon_{i1}\epsilon_{i2}}{\sum_{i=1}^N \epsilon_{i2}^2} = \frac{\lambda_1}{\lambda_2}.$$

With no measurement error ϵ_{i2} exactly equals $\lambda_2\omega_i$. By multiplying ϵ_{i2} with the estimate of $\frac{\lambda_1}{\lambda_2}$ one obtains a consistent estimator for $\lambda_1\omega_i$.

A second scenario is when outcomes from same equation are observed over time and the unobserved factor and its affect on the outcome are fixed over time. In this case, observations from different times provide “replicates”. The error structure is now given by

$$\epsilon_{is} = \lambda_1\omega_i + \varepsilon_{is}, \tag{11}$$

$$\epsilon_{it} = \lambda_1\omega_i + \varepsilon_{it} \tag{12}$$

where i indexes agents and 1 and 2 have been replaced with s and t , $s \neq t$, to emphasize the times series nature of the data. If the variance in the

errors is time-invariant, the simple average $\frac{1}{2}(\epsilon_{is} + \epsilon_{it})$ provides a minimum variance unbiased estimate of the unobserved factor. Note the importance of the replicate observation; without it, $\lambda_1\omega_i$ is not separately identified from ϵ_i . However, unlike the first case, where an exact estimate of $\lambda_1\omega_i$ is available, now only an unbiased estimate for $\lambda_1\omega_i$ is available. This case is analogous to estimating “fixed effects” directly.¹²

The last case in this subsection returns to the original case with two different equations for i . The question arises in the cross-section as to how one proceeds when the variance of ϵ_{i2} is not zero (the usual case). One can still proceed by regressing ϵ_{i1} on ϵ_{i2} , but in this case the least squares estimate of $\frac{\lambda_1}{\lambda_2}$ is not consistent. Specifically,

$$plim \frac{\sum_{i=1}^N \epsilon_{i1}\epsilon_{i2}}{\sum_{i=1}^N \epsilon_{i2}^2} = \frac{\lambda_1\lambda_2\sigma_\omega^2}{\lambda_2^2\sigma_\omega^2 + \sigma_{\epsilon_2}^2} = \frac{\lambda_1}{\lambda_2 + \sigma_{\epsilon_2}^2/\lambda_2\sigma_\omega^2}, \quad (13)$$

so the asymptotic bias in the estimate of the ratio $\frac{\lambda_1}{\lambda_2}$ is given by

$$\frac{-\lambda_1}{(\lambda_2^3\sigma_\omega^2/\sigma_{\epsilon_2}^2) + \lambda_2}.$$

The bias will be small when the variance in the unobserved factor is large relative to the variance in the measurement error. However, it is not possible to get rid of this bias without further information (like an additional equation).

This does not preclude one from using a simple average (for example) of ϵ_{i1} and $\frac{\hat{\lambda}_1}{\lambda_2}\epsilon_{i2}$ as an estimate for $\lambda_1\omega_i$. This estimate has bias for $\lambda_1\omega_i$ given by

$$1/2 * \frac{-\lambda_1\omega_i}{(\lambda_2^2\sigma_\omega^2/\sigma_{\epsilon_2}^2) + 1}.$$

Even in cases where the variance of the measurement error is not small, the biased estimate for $\lambda_1\omega_i$ may provide a significant improvement over assuming $\lambda_1\omega_i = 0$ for all i and/or $\lambda_1\omega_i = \lambda_1\omega_i + \epsilon_i$. Thus, unless it is believed that ϵ_2 is measured so poorly (for example) as to be completely useless in the analysis, the distribution of heterogeneity might be constructed using this method to see if the policy estimate is very sensitive to it; if not, one might feel a little more confident that the effects of unobserved heterogeneity are not so pernicious.

¹²Of course, the usual motivation for including fixed effects is to protect against correlation of ω_i with the endogenous Y 's.

4.3 Three Equations/One Unobserved Factor

When a third equation is added to the system of two equations/one unobserved factor, there will always be a minimum variance unbiased estimate available for the unobserved factor. Three observations with uncorrelated errors are available for it. To see this, let the three equation system be given by

$$\begin{aligned}\epsilon_{i1} &= \lambda_1\omega_i + \varepsilon_{i1} \\ \epsilon_{i2} &= \lambda_2\omega_i + \varepsilon_{i2}, \\ \epsilon_{i3} &= \lambda_3\omega_i + \varepsilon_{i3},\end{aligned}$$

with ϵ_{i3} the new residual from the third structural equation. Here the ratio of the covariance of ϵ_{i1} and ϵ_{i2} divided by the covariance of ϵ_{i3} and ϵ_{i2} equals $\frac{\lambda_1}{\lambda_3}$. Similarly, the ratio of the covariance of ϵ_{i1} and ϵ_{i3} divided by the covariance of ϵ_{i2} and ϵ_{i3} equals $\frac{\lambda_1}{\lambda_2}$. Finally the “unweighted” average

$$(\epsilon_{i1} + \frac{\lambda_1}{\lambda_2}\epsilon_{i2} + \frac{\lambda_1}{\lambda_3}\epsilon_{i3})/3 \tag{14}$$

provides an unbiased estimate for $\lambda_1\omega_i$. A similar approach can be used to recover $\lambda_2\omega_i$ and $\lambda_3\omega_i$. Optimal weighting is based on the differences in variance of each component in the averages are available.

With panel data, when the coefficient on the unobserved factor remains constant over time, not only are unbiased estimates for $\lambda_1\omega_i$ (for example) available, but the rate at which the variance in the error of the estimate falls is multiplicative in the number of time varying observations. For example, if there are two time series observations on each equation, there are six observations on $\lambda_1\omega_i$. With four time series observations, there are twelve observations on $\lambda_1\omega_i$. Thus, the availability of even a few replicate observations on any equation can substantially increase the precision of the unbiased estimate for $\lambda_1\omega_i$.

4.4 Two Unobserved Factors

In the unrestricted two-unobserved-factor case, every error ϵ_{ij} from the structural equations is a function of both unobserved factors $\omega_i = (\omega_{i1}, \omega_{i2})'$ and

a measurement-type error ε_{ij} . The unobserved factors' contribution to any equation j is given by

$$\lambda'_j \omega_i = \lambda_{j1} \omega_{i1} + \lambda_{j2} \omega_{i2}.$$

Without further restrictions, a minimum of six equations are required when two unobserved factors are present. When six equations are available, at least three unbiased observations (with uncorrelated errors) on $\lambda'_j \omega_i$ are available. With additional restrictions, the required number of equations for typically decreases).

The six equation setup is given as

$$\begin{aligned} \epsilon_{i1} &= \lambda_{11} \omega_{i1} + \lambda_{12} \omega_{i2} + \varepsilon_{i1} \\ \epsilon_{i2} &= \lambda_{21} \omega_{i1} + \lambda_{22} \omega_{i2} + \varepsilon_{i2} \\ \epsilon_{i3} &= \lambda_{31} \omega_{i1} + \lambda_{32} \omega_{i2} + \varepsilon_{i3} \\ \epsilon_{i4} &= \lambda_{41} \omega_{i1} + \lambda_{42} \omega_{i2} + \varepsilon_{i4} \\ \epsilon_{i5} &= \lambda_{51} \omega_{i1} + \lambda_{52} \omega_{i2} + \varepsilon_{i5} \\ \epsilon_{i6} &= \lambda_{61} \omega_{i1} + \lambda_{62} \omega_{i2} + \varepsilon_{i6}. \end{aligned}$$

Suppose a quantity of interest is $\lambda'_1 \omega_i$, the unobserved factors' effect on equation 1. Without loss of generality, the variance of ω_i is normalized so $E[\omega_i \omega_i'] = I$, the identity matrix.¹³ Let the matrix representation of the four factor loadings for equations one and two be given by

$$\Lambda_{12} = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix},$$

and similarly for Λ_{34} and Λ_{56} .

The approach is similar to that when there is one unobserved factor. Covariances between the errors from the simultaneous equations are used to construct consistent estimators for mappings that translate these errors $\epsilon_{ij}, j = 2, \dots, 6$ into random variables with mean $\lambda'_1 \omega_i$. These new random variables provide (noisy) observations on $\lambda'_1 \omega_i$; when averaged over they provide the unbiased, finite-variance estimator. The mappings are described next, followed by a discussion of their consistent estimators.

¹³The factor loadings (the λ 's) are identified only up to a right hand transformation.

Consider the 2×2 matrix given by

$$B = \Lambda_{12}\Lambda_{34}^{-1}.$$

When applied to the column vector that stacks ϵ_{i3} and ϵ_{i4} , one obtains

$$\begin{aligned} B \begin{pmatrix} \epsilon_{i3} \\ \epsilon_{i4} \end{pmatrix} &= \Lambda_{12}\Lambda_{34}^{-1} \left(\Lambda_{34}\omega_i + \begin{pmatrix} \epsilon_{i3} \\ \epsilon_{i4} \end{pmatrix} \right) \\ &= \Lambda_{12}\omega_i + \Lambda_{12}\Lambda_{34}^{-1} \begin{pmatrix} \epsilon_{i3} \\ \epsilon_{i4} \end{pmatrix}. \end{aligned}$$

Conditional on ω_i , the first element in this 2×1 column vector provides the second unbiased observation on $\lambda_1'\omega_i$ (the first is given by ϵ_{i1}). The error is a linear combination of ϵ_{i3} and ϵ_{i4} , and is thus uncorrelated with ϵ_{i1} , the error in ϵ_{i1} .

The third unbiased observation is similarly derived. Define the 2×2 matrix

$$G = \Lambda_{12}\Lambda_{56}^{-1}.$$

When applied to the column vector that stacks ϵ_{i5} and ϵ_{i6} , one obtains

$$G \begin{pmatrix} \epsilon_{i5} \\ \epsilon_{i6} \end{pmatrix} = \Lambda_{12}\omega_i + \Lambda_{12}\Lambda_{56}^{-1} \begin{pmatrix} \epsilon_{i5} \\ \epsilon_{i6} \end{pmatrix}.$$

The first element in this 2×1 column vector provides the third unbiased observation on $\lambda_1'\omega_i$. The error in this case is a linear combination of ϵ_{i5} and ϵ_{i6} , which is uncorrelated with the errors from both previous observations.

A consequence of the insights in Madansky (1964) is that consistent estimators for B and G are available as long as the 2×2 variance matrix for each set of two equations (e.g. $(\epsilon_{i3}, \epsilon_{i4})$) has full rank. In this case,

$$\begin{aligned} & \text{plim} \sum_{i=1}^n \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{pmatrix} \begin{pmatrix} \epsilon_{i5} \\ \epsilon_{i6} \end{pmatrix}' \left[\sum_{i=1}^n \begin{pmatrix} \epsilon_{i3} \\ \epsilon_{i4} \end{pmatrix} \begin{pmatrix} \epsilon_{i5} \\ \epsilon_{i6} \end{pmatrix}' \right]^{-1} \\ &= \Lambda_{12}E[\omega_i \omega_i']\Lambda_{56}' [\Lambda_{34}E[\omega_i \omega_i']\Lambda_{56}']^{-1} \\ &= \Lambda_{12}\Lambda_{56}'[\Lambda_{34}\Lambda_{56}']^{-1} \\ &= \Lambda_{12}\Lambda_{34}^{-1}. \end{aligned}$$

Similarly, a consistent estimate for G is given by

$$\begin{aligned}
 & plim \sum_{i=1}^n \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{pmatrix} \begin{pmatrix} \epsilon_{i3} \\ \epsilon_{i4} \end{pmatrix}' \left[\sum_{i=1}^n \begin{pmatrix} \epsilon_{i5} \\ \epsilon_{i6} \end{pmatrix} \begin{pmatrix} \epsilon_{i3} \\ \epsilon_{i4} \end{pmatrix}' \right]^{-1} \\
 &= \Lambda_{12} \Lambda'_{34} [\Lambda_{56} \Lambda'_{34}]^{-1} \\
 &= \Lambda_{12} \Lambda_{56}^{-1}.
 \end{aligned}$$

If additional restrictions across these six equations are available - for example, if one is a replicate observation of another - then fewer equations are generally required to achieve identification, and, given that identification of B and G has been achieved, more unbiased observations on the combined effect of the unobserved factors in any equation j are available.

4.5 The General Case

The logic of this approach extends to more general cases. See Madansky (1964).

5 Application: Consumer Demand

A consumer's valuation of a good that has both an adoption and a utilization dimension is reflected in the adoption decision and how much the good is used if adopted. These types of goods includes durables, which make up 30% of consumer expenditures, and products priced using two part tariffs, like (for example) land-line and cell phones. For these types of goods, this section shows how product utilization rates can help to account for otherwise unobserved heterogeneity in demand (and welfare).

The application uses data on television viewing to improve the estimates of demand curves for cable television. The basis of the demand estimation is the regression of market-level shares on prices, average consumer demographics in the market, and product characteristics of the available goods. The reduced form model uses a market-level representative agent demand model, the kind most popular in the demand literature (because market-level data is most readily available for practitioners). The structural equation conditions on the

market-level average of weekly hours watched of television, and is used to construct the reduced form equation that conditions on unobserved factors. Estimates of demand and welfare are compared across the linear, log-linear, log-log, and logit demand models.

5.1 Demand and Utilization Model

We begin with a generic reduced form demand equation, different variants of which will be considered for welfare estimation. This equation is given by

$$\tilde{Q} = \gamma Z + \gamma_P P + \eta, \quad (15)$$

where $\tilde{Q} = g(Q)$ is the transformation of Q that achieves linearity, Z is a vector of the observed product and household characteristics (and interactions) that are conditioned upon, γ is the vector of associated parameters, γ_P is the price coefficient, P is the cost of purchasing the good, and η is the reduced form error.

For goods that have a utilization aspect to them, the amount demanded of the good is determined simultaneously with the decision of how much the good will be used. The simultaneity of the decisions does not imply either one “causes” the other. Instead, both decisions are determined by observed and unobserved factors. To see how all of the underlying observed and unobserved exogenous factors enter into any reduced form demand equation - which is the equation used in utility theory - one must consider the structural equations for demand and use.

The structural equation for demand is given by

$$\tilde{Q} = Z_Q \theta_Q + \theta_{Q,P} P + \theta_{Q,U} U + \epsilon_Q, \quad (16)$$

where Z_Q is the vector of relevant observed product and household characteristics (and interactions) after conditioning on usage U , θ_Q is the vector of associated parameters, $\theta_{Q,P}$ ¹⁴ is the sensitivity of demand to the adoption price

¹⁴Generally, subscripts of the form (A, B) denote a coefficient from the structural equation with the left-hand side variable given by A (that is, with A 's coefficient normalized to one) and a right-hand side endogenous variable given by B .

conditional on use U , $\theta_{Q,U}$ gives the change in quantity demand if utilization increases one unit holding adoption price and Z_Q constant, and ϵ_Q reflects both unobserved demographics and idiosyncratic household tastes.

The structural equation for utilization is given by

$$U = Z_U \theta_U + \theta_{U,Q} \tilde{Q} + \theta_{U,P_U} P_U + \epsilon_U, \quad (17)$$

where θ_U is a vector of parameters that relates exogenous factors Z_U to use, $\theta_{U,Q}$ is the increase in use due to a one unit increase in demand, θ_{U,P_U} relates the amount of use undertaken to its unit price, and ϵ_U represents unobserved factors affecting use.¹⁵

We consider a two component specification for the error structure in the simultaneous equations. The first component, given by ω_i , is a single common factor entering each equation (with different coefficients) that represents the systematic component relating to both demand and use. The second component is an idiosyncratic i.i.d. shock that we want to separate from the systematic component for purposes of the demand and welfare calculations.¹⁶ These

¹⁵One can interpret these equations as the first order conditions for a representative agent model derived from a utility function given by $U^* = (Y_i - P_Q Q - P_U U) + V(Q, U, Z, \theta) + Q \epsilon_Q + U \epsilon_U$. Economists have a long history of estimating equations similar to (17). The parameters of interest are usually (the equivalent of) $\theta_{U,P}$ or $\theta_{U,Q}$, and the econometric discussion revolves around the fact that the demand decision (for appliances say) is endogenous (see Dubin and McFadden (1984)).

¹⁶One might consider a more general error structure with two unobserved factors given by the column vector $\omega_i = [\omega_{i1}, \omega_{i2}]'$. In addition, additive i.i.d. errors that do not count toward the systematic component of the demand curve are given by ϵ_{iQ} and ϵ_{iU} , both also mean zero. The errors decompose as

$$\epsilon_{iQ} = \lambda_Q \omega_i + \epsilon_{iQ} \quad (18)$$

$$\epsilon_{iU} = \lambda_U \omega_i + \epsilon_{iU}, \quad (19)$$

where $\lambda_Q = [\lambda_{Q1}, \lambda_{Q2}]$ and $\lambda_U = [\lambda_{U1}, \lambda_{U2}]$ are the relevant row vectors of parameters associated with the unobserved factors. For example, one unobserved factor might represent idiosyncratic taste for the good, or an omitted product characteristic (with market-level data). This factor would likely enter both the demand and use equation with positive coefficients (i.e. λ_{Q1} and λ_{U1} both positive). The second unobserved factor might represent “magnitude of time constraint”, which could enter demand and use equations with different signs; availability of the good is more important when the option value is high, but overall

errors are given by ϵ_{iQ} and ϵ_{iU} , both also mean zero. The errors decompose as

$$\epsilon_{iQ} = \lambda_Q \omega_i + \epsilon_{iQ} \quad (20)$$

$$\epsilon_{iU} = \lambda_U \omega_i + \epsilon_{iU}, \quad (21)$$

where λ_Q and λ_U determine how the common component enters each equation.

Equations (16) and (17) fully characterize the system of interest for many products at the center of policy/research questions. It is helpful to consider the relevant prices entering into each equation for some of these cases. For durable goods the adoption price is evident, and the price of use is given typically by the marginal cost of the energy consumed by the durable. In electricity demand and appliance choice models, the price of utilization is the price per hour of electricity times the amount of electricity per hour the appliance (e.g. air conditioner, heater, lighting, uses). In the vehicle demand literature, the price of use is the amount of miles the vehicle can be driven per dollar of gasoline. This econometric structure also arises for goods where users face a two part tariff (that is, subscription and use charges). For example, cell and land-line telephone services typically charge a fixed flat fee and/or require the purchase of a phone; this price enters the adoption equation. The price of use is then the marginal cost of using the phone an additional minute. Similarly, multi-channel video television like cable or satellite dish, the subscription price is the price for access. The marginal cost of viewing is the price of the time an individual spends watching television, that is, the opportunity cost of that hour for the individual.¹⁷

The reduced form parameters entering (15) are the statistics of interest for many questions because they do not condition on usage, that is, because the reduced form demand equation is *the* equation from utility theory. However, one drawback from estimating the reduced form directly is that the structural error from the demand equation and the utilization equation are confounded, and this prevents one from consistently estimating the systematic correlation

use is lower because there is less time to use it.

¹⁷Except for services like pay-per-view, there is no marginal charge for watching additional television.

of unobserved factors across equations.

Setting aside for the moment the endogeneity of price in the demand equation (which we address in the estimation), one strategy for identification of the parameters in the structural equations is available for many of the cases just described. It requires two exclusion restrictions to hold. First, conditional on use, the price of use must not enter the adoption equation. Similarly, the second restriction is that, conditional on adoption, the adoption price does not enter the use equation. These two restrictions are sufficient to identify all of the parameters in the structural equation setting.

By solving the structural equations for \tilde{Q} , the reduced form demand equation can be fully characterized in terms of the structural parameters, exogenous variables, and structural errors:¹⁸

$$\tilde{Q} = \frac{1}{1 - \theta_{Q,U} * \theta_{U,Q}} * [Z_Q \theta_Q + \theta_{Q,P} P + \theta_{Q,U} Z_U \theta_U + \theta_{Q,U} \theta_{U,P_U} P_U + \theta_{Q,U} \epsilon_U + \epsilon_Q]. \quad (22)$$

In terms of the (15), the observed factors are given by $Z = (Z_Q, P_U, Z_U)$, and γ are the associated parameters. For example, the coefficient on price – usually of particular importance for policy questions – is the function of structural parameters given by

$$\gamma_P = \frac{1}{1 - \theta_{Q,U} * \theta_{U,Q}} * \theta_{Q,P}, \quad (23)$$

where price sensitivity conditional on usage $\theta_{Q,P}$ is grossed up by $\frac{1}{1 - \theta_{Q,U} * \theta_{U,Q}}$ to account for the fact that adoption enters the use equation.¹⁹ Finally, the reduced form error is given by

$$\eta = \frac{1}{1 - \theta_{Q,U} * \theta_{U,Q}} * [\theta_{Q,U} \epsilon_U + \epsilon_Q]. \quad (24)$$

5.2 Welfare

The literature on demand and welfare estimation is largely dominated by four functional forms: linear, log-linear, log-log, and logit (or logit-like discrete

¹⁸The delta method or a bootstrap method can be used to compute standard errors for the reduced form parameters.

¹⁹More generally, to allow demographics to rotate the demand curve (in addition to shifting it) the price coefficient could be written as the function $\gamma_P(Z)$, the scalar value output of a multivariate function of consumer demographics and parameters.

choice models). Transformed back to levels, the quantity (here share) equation is given by

$$Q(P; Z, \omega, \theta, \lambda).$$

²⁰ The change in welfare for any agent that occurs when price increases from P_0 to P_1 is

$$\int_{P_0}^{P_1} Q(v; Z, \omega, \theta, \lambda) dv. \quad (25)$$

Summing over the distribution of (Z, ω) pairs yields the measure of aggregate welfare:

$$\int_{Z, \omega} \int_{P_0}^{P_1} Q(v, Z, \omega, \theta, \lambda) dv dP(Z, \omega). \quad (26)$$

For the applied researcher, $dP(Z, \omega)$ is not typically known because ω is not generally observed. The two “extreme” solutions that the literature adopts are described, followed by a third alternative we suggest.

5.2.1 Add None of the Error Back

As Hausman and Newey (1995) note, the log-log and log-linear demand frameworks almost always ignore the residual in the welfare calculation. The reason, in their words, is it is “difficult to improve without more information about the residual.” Thus, the measure used by practitioners for aggregate welfare in this case is given as

$$\int_Z \int_{P_0}^{P_1} Q(v, Z, 0, \theta) dv dP(Z), \quad (27)$$

with $\omega_i = 0$ for all i during the integration (no unobserved heterogeneity is permitted). Again, the non-linear manner in which ω_i enters (27) means this approach is inconsistent.

5.2.2 Add All of the Error Back

Another alternative is the other extreme: treat all of the demand error as demand-shifter/taste. This assumption is universally employed within discrete

²⁰For simplicity, income effects are not treated here, although they could be added back to the discussion at a cost to exposition.

choice frameworks, where consumer adoption decisions are modeled as being determined by the latent variable “utility”, a function of underlying tastes parameters and observed and unobserved product characteristics. In these frameworks, deviations between observed and predicted outcomes are entirely explained by differences in unobserved utility (or unobserved taste heterogeneity). Thus, *all* of the error is compensated in the welfare calculation, because it all represents utility.²¹

We use the logit demand model to provide an alternative along these lines in our results analysis.²² We also compute the log-log and log-linear models under this assumption. The welfare computation for these cases is given as

$$\int_{Z,\eta} \int_{P_0}^{P_1} Q(v, Z, \eta, \theta) dv dP(Z, \eta), \quad (28)$$

where $Q(\cdot)$ is the estimated reduced form demand equation retransformed to levels and η reflects the entire residual. If $\varepsilon_{iQ} \neq 0$, this approach is inconsistent, because the demand curve is shifting during integration due to measurement problems.

5.2.3 Add the Systematic Part of the Error Back

Our proposed alternative is to add the systematic part of the error back. The role of utilization data is to identify this systematic component. For example, if two individuals that look the same given the observed factors on them but differ substantially in their unobserved tastes, utilization data can help to more accurately recover the area under the true demand curve for each individual.

One approach is to regress ϵ_{iQ} on ϵ_{iU} (with no intercept). If the variance of ε_{iU} were zero, this regression would consistently estimate $\frac{\lambda_Q}{\lambda_U}$, that is,

$$plim \frac{\sum_{i=1}^N \epsilon_{iQ} \epsilon_{iU}}{\sum_{i=1}^N \epsilon_{iU}^2} = \frac{\lambda_Q \lambda_U \sigma_\omega^2}{\lambda_U^2 \sigma_\omega^2} = \frac{\lambda_Q}{\lambda_U}. \quad (29)$$

²¹This approach is probably taken because of the difficulties of dealing with measurement error in non-linear environments.

²²Using this model is also useful because it explicitly recognizes that the dependent variable only varies between zero and one. The logit demand model also has a very simple closed-form solution for welfare.

Multiplying $\epsilon_U = \lambda_U \omega_i$ by $\frac{\lambda_Q}{\lambda_U}$ yields $\lambda_Q \omega_i$ exactly, the shift in the demand curve that one needs for consistent welfare estimation. Thus, the exact aggregate welfare number given in (26) can be estimated consistently because the distribution of ω is estimated consistently.

Generally, the variance of ϵ_{iU} will not be zero. This means the regression will not produce a consistent estimate of $\frac{\lambda_Q}{\lambda_U}$. When a second observation on utilization is available, three independent, unbiased observations on $\lambda_Q \omega_i$ are available. This result is because

$$plim \frac{\sum_{i=1}^N \epsilon_{iQ} \epsilon_{iU_1}}{\sum_{i=1}^N \epsilon_{iU_2} \epsilon_{iU_1}} = \frac{\lambda_Q \lambda_{U_1} \sigma_\omega^2}{\lambda_{U_2} \lambda_{U_1} \sigma_\omega^2} = \frac{\lambda_Q}{\lambda_{U_2}}. \quad (30)$$

Multiplying ϵ_{iU_2} by the consistent estimator of $\frac{\lambda_U}{\lambda_{U_2}}$ yields a second unbiased estimate of $\lambda_Q \omega_i$. Similarly, a third unbiased estimate comes from substituting U_1 for U_2 and vice versa in the above calculation. An average of these three components can provide a significant improvement over the alternative approaches of either ignoring all of the error or adding all of it back. Weighting to account for differences in the variances can be particularly important, given the small number of observations on $\lambda_Q \omega_i$. The unbiased estimates of $\hat{\omega}_i$ are then used in the integration in place of 0 or η_i for all i , so demands use $\hat{\omega}_i$ as shifters, and welfare is estimated using

$$\int_{Z, \hat{\omega}} \int_{P_0}^{P_1} Q(v, Z, \hat{\omega}, \theta, \lambda) dv dP(Z, \hat{\omega}). \quad (31)$$

When a third observation on the underlying unobserved factor is not available, it is still possible to reduce the bias in demand estimation. The approach with just one demand and one use observation is as follows: regress ϵ_Q on ϵ_U , yielding

$$plim \frac{\sum_{i=1}^N \epsilon_{iQ} \epsilon_{iU}}{\sum_{i=1}^N \epsilon_{iU}^2} = \frac{\lambda_Q \lambda_U \sigma_\omega^2}{\lambda_U^2 \sigma_\omega^2 + \sigma_{\epsilon_U}^2} = \frac{\lambda_Q}{\lambda_U + \sigma_{\epsilon_U}^2 / \lambda_U \sigma_\omega^2}. \quad (32)$$

The asymptotic bias in the estimate of the ratio $\frac{\lambda_Q}{\lambda_U}$ is given by

$$\frac{-\lambda_Q}{(\lambda_U^3 \sigma_\omega^2 / \sigma_{\epsilon_U}^2) + \lambda_U}.$$

The bias will be small when the variance in the unobserved factor is large relative to the variance in ε_U . Thus, using a simple average (for example) of ϵ_{iQ} and $\frac{\hat{\lambda}_Q}{\lambda_U} \epsilon_{iU}$ as an estimate for $\lambda_Q \omega_i$ has bias for $\lambda_Q \omega_i$ given by

$$1/2 * \frac{-\lambda_Q \omega_i}{(\lambda_U^2 \sigma_\omega^2 / \sigma_{\varepsilon_U}^2) + 1}.$$

Even in cases where the variance of the measurement error is not small, the biased estimate for $\lambda_Q \omega_i$, when averaged with ϵ_Q , may provide a significant improvement over assuming $\omega_i = 0$ for all i or adding back ϵ_Q in its entirety. Thus, unless it is believed that the second instrument is measured with such a great amount of error as to be completely useless in the analysis, the policy estimate should be constructed both ways to see if it differs substantially; if not, one might feel a little more confident that the effects of unobserved heterogeneity are not so pernicious.

5.3 Data

The definition of goods in this market follows that in Goolsbee and Petrin (2004), with the overall set of goods approximated by the four main choices: antenna only, expanded basic cable, expanded basic cable plus some premium television (like the a la carte movie channel Home Box Office), and satellite dish. The focus of this paper will be on demand for expanded basic cable, which includes the local channels available via “antenna only” like ABC, NBC, CBS, Fox, and PBS, plus additional channels like ESPN (a sports channel), TNT (primarily movies), and others, with an average cable franchise providing 62 channels on expanded basic. Premium channels are available a la carte if and only if expanded basic cable has been adopted. Finally, satellite dish is a multi-channel video option like cable that has a relatively small market share of 11% in 2001.

The basis of estimation will be data on market-level shares and demographics along with the cable characteristics of expanded basic that households face in their cable franchise market.²³ This market-level data typifies the kind most

²³Each cable franchise is considered its own market, and almost all consumers in 2001 have one and only one cable company in their market.

often available to practitioners. Average demographics and market shares for expanded basic come from Forrester Technographics, 2001, a household-level survey which is meant to be nationally representative.²⁴ To minimize sampling error problems in market share estimates from less populated areas, the analysis is restricted to 265 cable franchise markets for which at least 30 respondents exist. For these markets, 70% subscribe to either expanded basic or expanded basic plus premium cable, which compares closely to the 68% reported as the aggregate cable share in Federal Communications Commission 01-389, the 2001 annual report on the status of competition in multichannel video markets. Table 1 reports summary statistics for all of the aggregates used in the analysis.

The exercise here will focus on estimating demand for expanded basic for two reasons. First, with an average market share of 47%, it is the most popular of the four choices. Second, any one premium option rarely costs more than \$10 per month, providing an upper bound on expanded basic welfare estimates that is not imposed during the estimation.

The utilization variable is derived from a Forrester survey question which asks *individuals* whether they (not the household) watch 0, 1-2, 3-5, 6-10, 11-15, or 16+ hours of television a week. Averaging these responses across households in every market (evaluating each bin except the largest at the midpoint) yields the variable television viewing hours (tvhours). This estimate is a lower bound on household television viewing because the top-coded bin is evaluated at 17 hours and the data is only reported for one individual in the household.²⁵ Except for the variables “all adults work” and “free time”, which are excluded from the demand equation conditional on utilization, all of the other variables listed in Table 1 enter the cable adoption equation.

Since television viewing is endogenous, an exclusion restriction is necessary to consistently identify its effect on the adoption decision. The excluded variables are constructed from survey answers to the questions “How many hours of free time (time which excludes work, chores, errands) do you have

²⁴More details about it can be found in Goolsbee and Petrin (2004).

²⁵The exact underlining and question is “How may hours per week do you spend watching TV?”.

Table 1
Summary Statistics: Market Shares and Household Averages

Variable	Mean	Std. Dev.
Expanded Basic*	0.47	0.11
Expanded Basic + Premium*	0.23	0.10
Television Watched (Hrs. Weekly)	8.07	0.81
Free Time (Hrs. Weekly)	20.17	2.04
All adults work*	0.29	0.08
Married*	0.69	0.10
Household Size	2.61	1.23
Household Income	\$56.0K	\$10.5K
Household Assets	\$220.8K	\$92.5K
Observations	265	

Source: Forrester Technographics, 2001.

* designates a share variable.

per week, including weekend hours?” and “Do all adults in the household work full time?” Thus, the identification assumption is that these variables, which affect the amount of time available to watch television, only affect cable television adoption decisions through their effect on television viewing. In the data television viewing across markets increases with increases in free time, and decreases with increases in the fraction of households that have both adults working.

Table 2 summarizes the prices and characteristics of cable companies used to estimate demands. Consumer preferences can be affected by the channel capacity of a cable system, the price of expanded basic for the system, the price of expanded basic plus premium, the number of premium channels available, and the number of over the air channels available in the market area. The Factbook also provides the local franchise tax as a percent of revenue paid by the cable company to the local community.

Since prices for antenna and satellite dish do not vary across markets, only

Table 2
Summary Statistics: Television Markets

Variable	Mean	Std. Dev.
Monthly Expanded Basic Price	\$27.17	\$5.68
Monthly HBO Price	\$11.44	\$1.38
Channel Capacity	66.13	22.46
Premium Channels	5.56	1.45
Over-Air Channels	10.91	4.39
City Fixed Fee (percent)	4.29	1.14
Observations	265	

Source: Warren Publishing Television and Cable Factbook, 2001.

the prices of expanded basic and expanded basic plus some premium enter the demand equation.²⁶ There is substantial price variation for the monthly cost of expanded basic around the mean price of \$27.10, ranging from a minimum of \$15 to a maximum of \$45. In the data respondents only answer whether they have consumed some premium television. Expanded basic price plus the price of the most popular premium channel, Home Box Office (HBO), is used as the premium price proxy; consumers are paying at least this much to consume the most popular premium television option.

One concern when estimating a demand curve using this variation directly is the potential for price endogeneity.²⁷ Three instruments for the two endogenous prices are used. The first is the local franchise tax paid by the cable company to the local community. This is a percent of gross revenue that varies by market and is reported in Warren Publishing (2001). It is positively correlated

²⁶This presents no problems for estimating cable demands, but does pose problems for satellite demand estimates (see Goolsbee and Petrin (2004)).

²⁷If there are some characteristics of the local cable franchise that are known by both the consumers and the suppliers and if cable prices respond to these factors, the price elasticity will typically be biased toward zero. For example, a cable system with the same observables but with relatively good service will tend to be more desirable and have higher prices, making it seem as though consumer demand does not respond to high prices. Goolsbee and Petrin (2004) show this problem is a pronounced one for cable markets.

with prices. It is also not correlated with any of the observed characteristics, suggesting it may be uncorrelated with any unobserved characteristics that lead to price endogeneity problems (like customer service).

The last two instruments follow Hausman (1997) and Crawford (2001), averaging over the prices of the cable companies with the same multiple system operators (MSOs) but operating in different markets. These average prices reflect common cost side factors like programming costs shared by companies owned by the same MSO.²⁸ These instruments should exclude some of the idiosyncratic features of demand in each market that are correlated with prices in the market (like service).

5.4 Results

Following Hausman and Newey (1995), the focus begins with results from the log-log model, perhaps the most popular model used in demand estimation. Table 3 presents three different sets of demand estimates for the log-log model. The explanatory variables that enter the demand equation include channel capacity of the system, the number of premium channels available, the number of over-the-air channels, and market-level averages of household size, an indicator for marital status, and four indicator variables for five income groups and six indicator variables for seven wealth groups. To account for differences in sampling error, each market-level observation is weighted by the number of households observed in that market from the Forrester data.

The first column contains the OLS results without television hours entering the regression (i.e. the reduced form equation). Both price coefficients are very close to zero, which is consistent with the existence of unobserved characteristics like service that are positively correlated with price. The point estimates imply an elasticity of -0.39 for expanded basic cable, suggesting either an omitted variables problem or many cable monopolists pricing on the very inelastic part of the demand curve.

Column two presents the two stage least squares (2SLS) results for the no-

²⁸Larger MSOs with the ability to reach more advertising markets typically receive lower programming prices.

Table 3
Expanded Basic Demand Estimates
 Dependent Variable: Expanded Basic Market Share

	OLS	2SLS	2SLS
	Coefficient	Coefficient	Coefficient
	(Std. Error)	(Std. Error)	(Std. Error)
expanded basic price	-0.396 (0.299)	-1.911 (1.146)	-1.222 (1.198)
premium price	0.678 (0.399)	2.899 (1.575)	1.717 (1.677)
tvhours			0.137 (0.062)
channel capacity	-0.0002 (0.0006)	-0.0001 (0.0007)	0.0001 (0.0008)
premium channels	-0.043 (0.010)	-0.038 (0.011)	-0.042 (0.011)
over the air channels	-0.007 (0.003)	-0.005 (0.004)	-0.010 (0.004)
household size	-0.273 (0.088)	-0.265 (0.094)	-0.067 (0.130)
married	1.444 (0.243)	1.459 (0.259)	1.198 (0.287)
constant	-2.257 (0.635)	-5.487 (2.257)	-4.729 (2.303)
Root MSError	0.225	0.240	0.242
Observations	265	265	265

Note: All regressions include extensive controls for income and wealth, including four indicator variables for five income groups and six indicator variables for seven wealth groups. All of the income indicators and four of the six wealth indicators enter with t-statistics of at least one; six are significant at 5%. Income is negatively correlated with cable adoption conditional on wealth and wealth is positively correlated with cable adoption conditional on income.

utilization (reduced form) equation. The instruments for price are the market franchise tax and the average prices of same-MSO-but-different-market cable franchises (for both expanded basic and expanded basic plus premium). Both price coefficients increase by a factor of six. At the new point estimates, the aggregate elasticity for expanded basic cable is -1.91, comparable to previous expanded basic cable estimates found in the literature from Hazlett and Spitzer (1997), U.S. General Accounting Office (2000), Crawford (2000), and Goolsbee and Petrin (2004). Over-the-air channels and premium channels enter with the expected sign; increases in each are likely to increase antenna only and expanded basic plus premium shares respectively at the expense of expanded basic shares. There is one overidentification restriction that can be tested, and the validity of instruments/model is not rejected at the usual levels of significance (p-value=0.22). Finally, while the difference in elasticity estimates between the OLS and 2SLS results is 1.52 – large in economic terms – a Hausman test for a significant difference cannot reject the null of no difference (p-value=0.17).

Column three contains the 2SLS results conditional on utilization, with amount of free time and whether both adults work as instruments for the endogenous regressor tvhours. The unconditional price elasticity of demand implied by these conditional estimates is -2.17, similar to that obtained without utilization. The coefficient on market television viewing is significant at 5%, rejecting the demand specification without utilization. Conditional on expanded basic and expanded basic plus premium prices, a one hour increase in market television viewing (approximately one standard deviation) leads to an increase in the market share of expanded basic equal to 13%. In the levels, this translates into an increase in share of 6% on the base share of 47%.

Table 4 compares point estimates and standard errors, implied elasticities, and p-values for two overidentification tests, one for each equation. For every functional form television hours is significant at the 5% level. All of the estimates imply an elasticity close to 2. In column three Test 1 asks whether the two overidentification restrictions for the cable demand equation can reject the specification. The lowest p-value is 0.22, so no rejection of the null of correct

Table 4
**TVhours and Cable Elasticity Estimates, and
Two Overidentification Tests**

Method	TVhours Estimate (Std. Error)	Implied Cable Price Elasticity	P-Value for Over-ID Test 1	P-Value for Over-ID Test 2
Log-Log	0.13 (0.06)	-2.17	0.22	0.78
Log	0.13 (0.06)	-2.38	0.25	0.78
Logit	0.22 (0.11)	-2.26	0.33	0.79
Linear	0.53 (0.26)	-2.10	0.36	0.80

specification/valid instruments obtains at any meaningful level of significance. The second overidentification test in column four asks if the tvhours equation is well-specified. Again, no rejection obtains at any meaningful level of significance. The welfare implications of these estimates are considered next.

The average welfare change is equal to the average amount of household surplus in each market for expanded basic subscribers. Alternatively, for adopters of expanded basic in a market, it is the average difference in the reservation price minus the existing price of expanded basic, where a household's reservation price is defined as the price at which it substitutes to one of the other television products: expanded basic plus premium, satellite dish, or antenna only.

Cable monopolies' bundling of expanded basic with premium implies a specification test that has power to uncover overstatements of welfare, like those that often arise when one extrapolates the surplus measure to regions outside the observed price variation. For the surplus question the price change of expanded basic increases to ∞ with all other product prices held constant. Since expanded basic is bundled with premium in the expanded basic plus premium choice, surplus for expanded basic cannot, in theory, exceed the price for the a la carte premium channel; no rational consumer would purchase expanded basic when they could get expanded basic *plus* premium for a lower

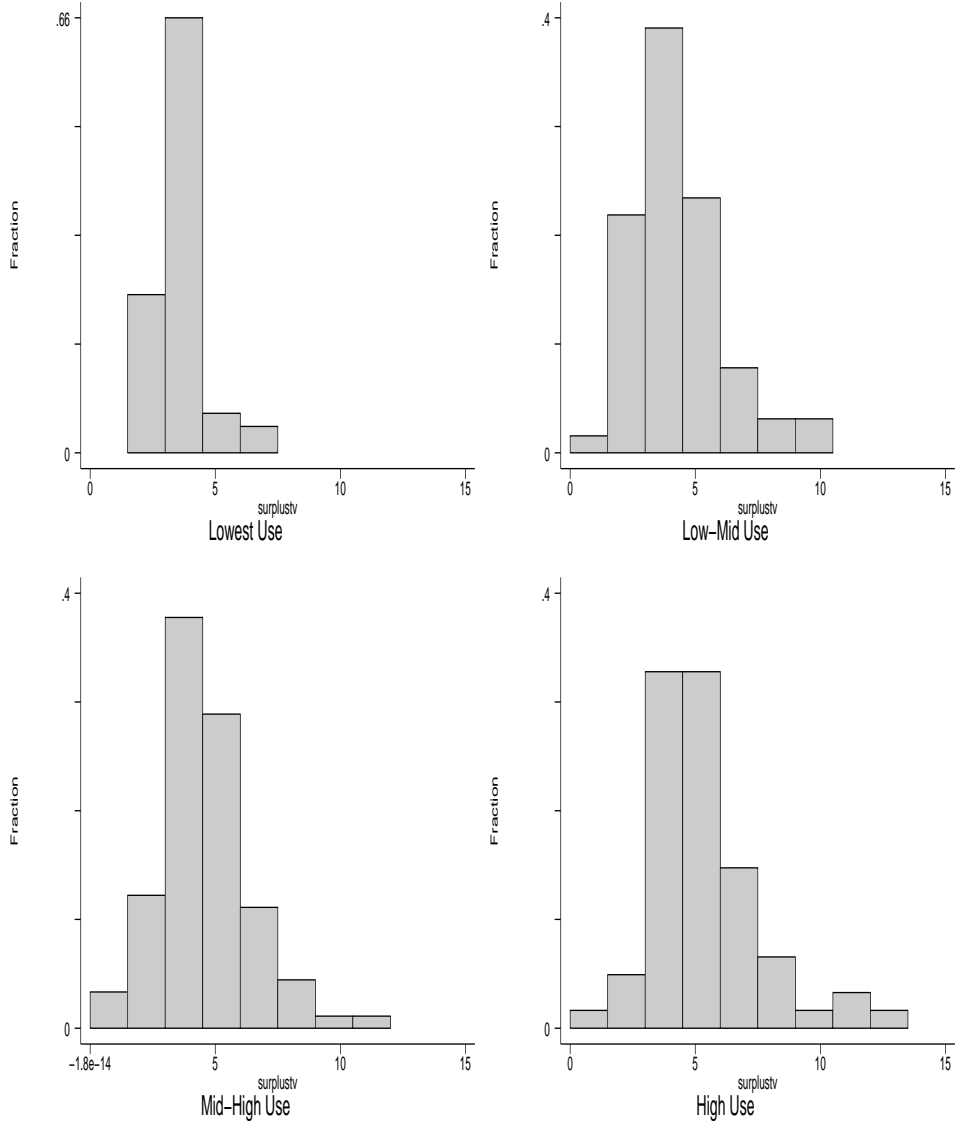
price. This is not imposed during the estimation of welfare.

The initial analysis focuses on the region of observed price variation, beginning with the monthly welfare change for the price increase to the maximum expanded basic cable price observed in the data. This is a best case scenario for the no use case using the most popular functional form. Percentage changes in surplus associated with using utilization relative to not using it are considered first. The median percentage change in the log-log case is a decrease in welfare of 8.5% relative to no use data; ignoring use tends on average to overstate welfare. The inframarginal markets are starkly affected by the use data; 5th and 95th percentiles of this distribution of percentage changes are equivalent to -38.3% and 50.2% respectively. Thus, without the use data, low use markets have their welfare substantially overstated, while high use markets have their welfare substantially understated.

Monthly welfare estimates for the log-log specification with the tvhours regressor are reported in Figure 1. The distribution of welfare for each of four television viewing groups is shown: lowest use, low-mid use, mid-high use, and high use. Only two of 265 markets have an average surplus estimate that exceeds the price of HBO in the market, suggesting the log-log specification using only observed price variation is a reasonable one. The economic significance of including television viewing in the cable demand equation is apparent; as viewing increases, the distribution of surplus shifts to the right, from a mean of \$3.63 per month for the low use group to a mean of \$5.46 per month for the high use group. These differences are statistically significant. Thus, the answer to any question related to the *distribution* of welfare or to infra-marginal markets improves with utilization data.

Figure 2 plots the distribution of absolute percentage changes in surplus associated with using utilization relative to not using it. The median absolute percentage change in the log-log case is 20.2% relative to no use data, with the 90th percentiles equal to 50.2% and the maximum equal to 120%. In only 25% of markets are these absolute percentage changes less than 10%, suggesting very different welfare implications for most markets once utilization is conditioned on.

Figure 1: Average Monthly Surplus: Expanded Basic Adopters



Surplus For Expanded Basic Using TV Hours: Log-Log

Figure 2: Surplus Changes With Use Data

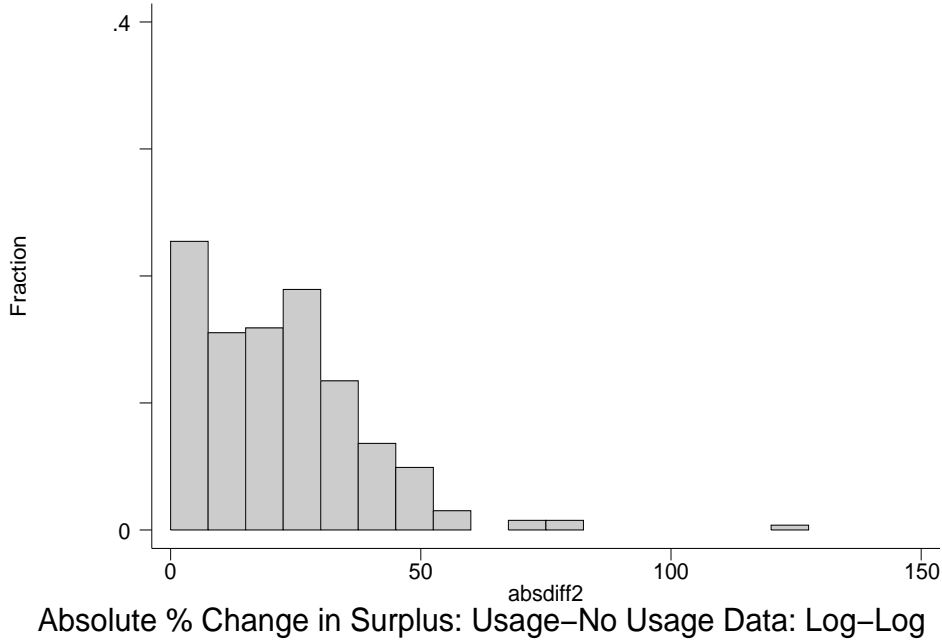


Table 5 summarizes results for all of the functional forms under consideration. The table has estimates for median surplus, the percent of markets failing the welfare specification test, and the median percentage welfare increases relative to not incorporating use data. Welfare estimates for the same specifications but for price increases only up to \$44 a month, the largest price observed in the data for expanded basic, are reported in the top half of Table 5. These numbers provide a lower bound on welfare that does not suffer from the extrapolating outside the region of observed price variation. Using this variation, the surplus numbers across specifications are similar, ranging from the minimum linear case of \$0.95 to the maximum log-log case of \$4.28. Only two markets of the 265 fail the specification test, and only for the log-linear and log-log case. Even for this modest price increase, median percentage decreases range from 8% to 75%.

The bottom part of the table contains results for price increases to the reservation price. Median monthly surplus estimates vary across the estimators

from \$2.89 for the linear case to \$9.05 for the log-log case. Welfare estimates using log-log demands exceed the price of HBO in almost 30% of markets, suggesting that functional form outside the region of observed price variation is driving these numbers. Median welfare estimates decrease between 11% and 76% over the no use case. When the distribution of absolute percentage changes is considered, the median change ranges from 23% to 76%. Overall, the use data plays an important role in the welfare estimates, and the results suggest that households are realizing perhaps \$3-\$5 per month in surplus with their expanded basic choice, with the higher use households receiving more surplus.

6 Conclusions

Many policy/welfare questions are non-linear functions of the observables and errors entering the model. In these cases, all the factors must be accounted for when the policy analysis takes place, because errors do not “average out.” Thus, from the applied person’s perspective, the problem is that relevant heterogeneity is not observed, and thus cannot be conditioned on during the analysis, leading to biased and inconsistent policy estimates.

This paper describes this bias and provides methods to alleviate it. The theme of the methods is that jointly determined variables contain information on underlying and unobserved factors. The implication - which has many applications - is that the errors from systems of related equations can be used to construct estimates of unobserved factors. A number of different cases common to empirical work are analyzed.

The application in the paper employs product utilization rates to account for otherwise unobserved taste heterogeneity when estimating demand and/or welfare. The utilization corrections are a consistency issue because unobserved taste heterogeneity does not enter the demand equation linearly, so these taste errors do not average out in the welfare calculation. As Hausman and Newey (1995) note, ignoring unobserved taste heterogeneity is “consistent with current practice in applied econometrics, and is difficult to improve without more

Table 5
**Median Surplus and Percentage Changes in Surplus
 (Use-NoUse)/Use
 for Price Increases to...**

...\$44, the Maximum Expanded Basic Price:				
Method	Median Surplus	% Surplus > P_{HBO}	Median Change %	Abs(Median) Change %
Log-Log	\$4.28	0.7%	-08.5%	20.2%
Log-Linear	\$3.88	0.7%	-11.6%	19.9%
Logit	\$0.95	0.0%	-75.7%	75.7%
Linear	\$2.11	0.0%	-20.2%	33.1%
...Estimated Reservation Prices:				
Method	Median Surplus	% Surplus > P_{HBO}	Median % Change	Abs(Median) % Change
Log-Log	\$9.05	27.9%	-29.9%	31.3%
Log-Linear	\$5.11	1.5%	-21.2%	25.0%
Logit	\$1.02	0.0%	-76.9%	76.9%
Linear	\$2.89	0.0%	-11.9%	23.6%

information about the residual.” The role of utilization data is precisely to provide this information; if observed, it can be used to condition out some of the unobserved taste heterogeneity that would otherwise be attributed to measurement error (and not conditioned on in the welfare calculation).

The empirical application in this paper uses data from the telecommunications industry. Variability in the amount of television watched is used to improve demand estimates for the adoption of cable television. To show the relevance of the methods for an applied case, the application uses market-level data, the most commonly available type of data, to evaluate the change in welfare from a large price increase in the expanded basic cable price. In particular, market-level averages of television viewing are shown to have, conditional on other market-level observables, significant explanatory power in the market-level adoption equation for cable television. For the linear, log-linear, log-log, and logit demand models, the distribution of welfare changes induced by a large price increase differs substantially when the utilization information is added. The median welfare change decreases by on average almost 20%. For some infra-marginal markets, the welfare change is misstated by between 50% and 80%.

References

- CRAWFORD, G. (2000): “The Impact of the 1992 Cable Act on Household Demand and Welfare,” *RAND Journal of Economics*, 31(3), 422–449.
- CRAWFORD, G. (2001): “The Discriminatory Incentives to Bundle: The Case of Cable Television,” Working Paper.
- DUBIN, J. A., AND D. L. MCFADDEN (1984): “An Econometric Analysis of Residential Electric Appliance Holdings and Consumption,” *Econometrica*, 52(2), 345–62.
- GOOLSBEE, A., AND A. PETRIN (2004): “The Consumer Gains from Direct Broadcast Satellites and the Competition with Cable Television,” *forthcoming Econometrica*.
- GRILICHES, Z. (1977): “Estimating the Returns to Schooling: Some Econometric Problems,” *Econometrica*, 45(1), 1–22.
- HANEMANN, W. M. (1984): “Discrete/Continuous Models of Consumer Demand,” *Econometrica*, 52(3), 541–62.
- HAUSMAN, J. (1979): “Individual Discount Rates and the Purchase and Utilization of Energy-Using Durables,” *The Bell Journal of Economics*, 10(1), 33–54.
- (1997): “Valuation of New Goods Under Perfect and Imperfect Competition,” in *The Economics of New Goods*, ed. by R. Gordon, and T. Bresnahan. Chicago: University of Chicago Press, chap. 5, pp. 209–237.
- HAUSMAN, J. A., AND W. K. NEWEY (1995): “Nonparametric Estimation of Exact Consumers Surplus and Deadweight Loss,” *Econometrica*, 63(6), 1445–76.
- HAZLETT, T., AND M. SPITZER (1997): *Public Policy Toward Cable Television*. The MIT Press and the AEI Press (Cambridge, Massachusetts and Washington, D.C.).

- HECKMAN, J., AND R. ROBB (1986): *Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes, in Drawing Inferences from Self-Selected Samples* Springer-Verlag, chap. 1, pp. 63–113.
- HECKMAN, J., AND J. SCHEINKMAN (1987): “The Importance of Bundling in a Gorman-Lancaster Model of Earnings,” *The Review of Economic Studies*, 54(2), 243–255.
- KRUEGER, A., E. HANUSHEK, AND J. K. RICE (2002): *The Class Size Debate*. Economic Policy Institute.
- MADANSKY, A. (1964): “Instrumental Variables in Factor Analysis,” *Psychometrika*, 29(2), 105–113.
- MANNERING, F., AND C. WINSTON (1985): “A Dynamic Empirical Analysis of Household Vehicle Ownership and Utilization,” *RAND Journal of Economics*, 16(2), 215–236.
- MINCER, J. (1974): *Schooling, Experience, and Earnings*. New York: NBER.
- NEYMAN, J., AND E. SCOTT (1948): “Consistent Estimates Based on Partially Consistent Observations,” *Econometrica*, 16(1), 1–32.
- PUDNEY, S. (1981): “Instrumental Variable Estimation of a Characteristics Model of Demand,” *The Review of Economic Studies*, 48(3), 417–433.
- (1982): “Estimating Latent Variable Systems When Specification is Uncertain: Generalized Component Analysis and the Eliminant Method,” *Journal of the American Statistical Association*, 77(380), 883–889.
- U.S. GENERAL ACCOUNTING OFFICE (2000): “The Effect of Competition from Satellite Providers on Cable Rates,” *Report to Congressional Requesters*, (GAO/RCED-00-164).
- WARREN PUBLISHING (2001): *Warren Publishing Television and Cable Factbook*. Warren Publishing Inc (Washington, D.C.).