Scanner Data and the Estimation of Demand Parameters

Richard Green, Zuhair A. Hassan, and S. R. Johnson

Working Paper 91-WP 73
April 1991
SCANNER DATA AND THE ESTIMATION OF DEMAND PARAMETERS

Approaches to the specification of demand systems and the estimation of demand parameters have expanded considerably during recent years (Blundell, 1988). Available empirical estimates have been based on time-series, cross-section, and time-series of cross-section data on individuals, households and more aggregated units. Studies by Capps and Nayga (1990), Capps (1989), Funk, Meilke, and Huff (1977), and Marion and Walker (1978) have employed scanner/retail sales data, a potentially rich new source of information for estimating demand elasticities. Specifically, Capps, and Funk, Meilke and Huff have utilized data from scanners/retail sales to estimate retail demand functions for meat products in the United States and Canada, respectively.

In these two scanner data studies, income (total expenditures) was omitted from the demand function specifications. Results reported indicate that the estimated own-price elasticities for the commodities are generally smaller in absolute terms than the corresponding cross-price elasticities, an unusual empirical finding. The objectives of this article are to consider potential problems that can arise using scanner data. These problems relate to the size of the estimated price elasticities, the consequences of selected specification errors and the efficiency impacts of the omission of variables. Clear understanding of the implications of these problem areas may prove useful for future users of scanner data in demand analysis.

Price Elasticities

The relationship found by Capps, and Funk, Meilke and Huff among the own-price and cross-price elasticities was in part due to the functional form that was used and the omission of an income variable. It is significant to
note that in the Capps and Nayga study, a proxy for income was included and the unusual relationship among the own- and cross-price elasticities was not found.

Consumer demand theory does not require that the own-price elasticity (in absolute terms) exceed the magnitudes of all of the individual cross-price elasticities for a particular commodity. However, under certain conditions it seems reasonable that this condition should hold empirically. Assume the commodities in question are (a) superior (positive income elasticities), (b) gross substitutes (positive uncompensated cross-price elasticities) or (c) independent (zero cross-price elasticities). These are plausible possibilities for the commodities analyzed by Capps, and Funk, Meilke and Huff. In addition, recall from the homogeneity condition that the sum of the own-price, cross-price, and income elasticities is zero (in other words, the sum of cross-price elasticities is equal to the difference between the own-price and income elasticities).

Under the above conditions and with homogeneity the own-price elasticity of the particular commodity group must be negative and larger (in absolute terms) than the sum of the cross-price elasticities. Thus, for normal (i.e., substitute or independent) and superior commodities, the absolute size of own-price elasticities will be larger than each of the individual cross-price elasticities (Tomek and Robinson, p. 54) in the demand equation.

Bias and Omission of Variables

To better understand the own-price and cross-price elasticities results from the above mentioned studies, the compound effects of the functional form and omitting income are evaluated. First, for the double logarithmic functional form used for the demand functions (Capps), homogeneity can hold
even though the symmetry and adding-up restrictions are not satisfied (Deaton
and Muellbauer, p. 17). Thus, homogeneity can be used to interpret the
empirical results. Second, income (total expenditure) was not included in the
demand relationships, although at the level of aggregation used, per capita
total expenditures of the meat subgroup could have been obtained from the
scanner data by assuming that consumers purchased all their meat at the same
chain of stores.

What is the impact of this specification problem on the relationship
between own-price and cross-price elasticities? Given the high degree of
correlation among the explanatory variables in the studies by Capps, and Funk,
Meilke and Huff, we would expect that the direction of the biases for
Seemingly Unrelated Regressions (SUR) estimates would not differ from the OLS
case when a relevant variable is omitted (Kmenta, pp. 443-446). Indeed, this
result has been shown to hold, under certain conditions, when a relevant
explanatory variable is omitted from a simple SUR model.

Using a suitable transformation for the SUR model the bias terms can be
derived specifically (Bacon 1974; Kmenta 1986). The transformation is
standard and allows the use of ordinary least squares estimation techniques
for the SUR model. Suppose for illustration that the true demand model is

\[ y_1 = X_1 \beta_1 + X_3 \beta_3 + \epsilon_1 \]  \hspace{1cm} (1)

\[ y_2 = X_2 \beta_2 + X_4 \beta_4 + \epsilon_2 \]  \hspace{1cm} (2)

where \( y_1 \) is a \( T \times 1 \) vector of observations on the dependent variable, \( X_i \) is a
\( T \times K_i \) matrix of values of the explanatory variables, \( \beta_i \) is a \( K_i \times 1 \) vector of
parameters, and \( \epsilon_i \) is a \( T \times 1 \) vector of disturbances, \( i = 1,2 \). The standard assumptions are

\[
E(\epsilon_i) = 0; \quad i=1,2 \quad \text{and} \quad \quad \quad \quad (3)
\]

\[
E(\epsilon_i \epsilon_j') = \sigma_{ij} I; \quad i,j = 1,2.
\quad \quad \quad \quad (4)
\]

From equation (4), the error terms across equations are contemporaneously correlated and there is no serial correlation across \( t \in T \).

Now suppose that \( X_3 \) and \( X_4 \) are omitted from the "true" model in (1) and (2). The disturbance terms become

\[
\epsilon_i^* = X_3 \beta_3 + \epsilon_i \quad \quad \quad \quad (5)
\]

and

\[
\epsilon_2^* = X_4 \beta_4 + \epsilon_2 \quad \quad \quad \quad (6)
\]

The SUR estimators of \( \beta_1 \) and \( \beta_2 \) can be obtained by applying a transformation due to Bacon and presented in Kmenta, pp. 640-41. The transformed SUR regressions are

\[
\begin{bmatrix}
I & 0 \\
\bar{a}_1I & \bar{a}_2I
\end{bmatrix} 
\begin{bmatrix}
Y_1 \\
Y_2
\end{bmatrix} = 
\begin{bmatrix}
I & 0 \\
\bar{a}_1I & \bar{a}_2I
\end{bmatrix} 
\begin{bmatrix}
X_1 & 0 \\
X_2 & \beta_2
\end{bmatrix} + 
\begin{bmatrix}
I & 0 \\
\bar{a}_1I & \bar{a}_2I
\end{bmatrix} 
\begin{bmatrix}
\epsilon_1^* \\
\epsilon_2^*
\end{bmatrix} \quad \quad \quad \quad (7)
\]
where \( a_1 = \pm \sqrt{\frac{\rho^2}{1-\rho^2}} \) and \( a_2 = \pm \sqrt{\frac{\sigma_{12}}{\sigma_{22}(1-\rho^2)}} \) and \( \rho = \sigma_{11}/\sqrt{\sigma_{11}\sigma_{22}} \). With the transformation (7), the new disturbance terms have variance-covariance matrix \( \sigma^2 I \), where \( \sigma^2 \) is a constant. Assume that \( a_1 \) and \( a_2 \) are known, i.e., that the contemporaneous covariances for the original disturbances are known. Then the SUR estimators of \( \beta_1 \) and \( \beta_2 \) from (7) are

\[
\begin{pmatrix}
 b_1 \\
 b_2
\end{pmatrix} = \left( \begin{bmatrix} X_1 & 0 \\ a_1X_1 & a_2X_2 \end{bmatrix} \begin{bmatrix} X_1 & 0 \\ a_1X_1 & a_2X_2 \end{bmatrix} \right)^{-1} \begin{bmatrix} X_1 & 0 \\ a_1X_1 & a_2X_2 \end{bmatrix} \begin{bmatrix} Y_1 \\ a_1y_1 + a_2y_2 \end{bmatrix} .
\] (8)

Taking expectations of both sides of (8) yields

\[
E(b) = (X'X)^{-1} X'E \begin{bmatrix} Y_1 \\ a_1y_1 + a_2y_2 \end{bmatrix}
\] (9)

where \( b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \) and \( X = \begin{bmatrix} X_1 & 0 \\ a_1X_1 & a_2X_2 \end{bmatrix} \). Since, the "true" values of \( y_1 \) and \( y_2 \) are given in (1) and (2),

\[
E(Y_1) = X_1\beta_1 + X_1\beta_2 \quad \text{and}
\] (10)
\[ E(y_2) = X_2\beta_2 + X_4\beta_4. \]  

Thus,

\[ E(b) = (X'X)^{-1}X' \left( X_1\beta_1 + X_2\beta_2 + a_1(X_1\beta_1 + X_3\beta_3) + a_2(X_2\beta_2 + X_4\beta_4) \right) \]

\[ = (X'X)^{-1}X' \left\{ X \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{bmatrix} X_1 \\ a_1X_3 \\ a_2X_4 \end{bmatrix} \begin{pmatrix} \beta_3 \\ \beta_4 \end{pmatrix} \right\} \]

\[ = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + (X'X)^{-1}X' \begin{pmatrix} X_2 \\ a_1X_3 \\ a_2X_4 \end{pmatrix} \begin{pmatrix} \beta_3 \\ \beta_4 \end{pmatrix}. \]

The bias is given by the second term on the right hand side of equation (12). After simplifying, the bias term becomes

\[ \begin{bmatrix} (1 + a_1^2)X_1X_1 & a_1a_2X_1X_2 \\ a_1a_2X_1X_2 & a_2^2X_2X_2 \end{bmatrix}^{-1} \begin{bmatrix} (1 + a_1^2)X_1X_3 \\ a_1a_2X_1X_4 \end{bmatrix} \begin{pmatrix} \beta_3 \\ \beta_4 \end{pmatrix}. \]

**Interpretation of Bias**

In general, it is obvious that when relevant variables are omitted the qualitative bias is difficult to determine. However, in special cases the sign of the bias can be obtained. First, let \( X_1, X_2, X_3, \) and \( X_4 \) be \( T \times 1 \)
vectors of observations on the explanatory variables. Furthermore, let $X_3 = X_4$, i.e., assume the same explanatory variable is omitted from each equation. With these assumptions, $X_1 \dot{X}_1 = \Sigma X_{1e}$, $X_2 \dot{X}_2 = \Sigma X_{2e}$, etc. Thus, after simplification the bias term is,

$$
\frac{1}{D} \begin{bmatrix}
(1 + a_1^2) a_2^2 \Sigma x_{1e}^2 \Sigma x_{2e}^2 - (a_1 a_2)^2 (\Sigma x_{1e}^2 \Sigma x_{2e}^2) \\
(1 + a_1^2) (a_1 a_2) (\Sigma x_{1e} x_{2e}) (\Sigma x_{1e} x_{3e}) + a_2 (1 + a_1^2) (\Sigma x_{1e}^2) (\Sigma x_{2e} x_{3e}) \\
 a_1^2 a_2 (\Sigma x_{2e}^2) (\Sigma x_{1e} x_{3e}) - a_1 a_2 (\Sigma x_{1e} x_{2e}) (\Sigma x_{2e} x_{3e}) \\
(1 + a_1^2) (a_1 a_2) (\Sigma x_{1e} x_{2e}) (\Sigma x_{1e} x_{3e}) + a_2 (1 + a_1^2) (\Sigma x_{2e}^2) (\Sigma x_{2e} x_{3e})
\end{bmatrix}
$$

$$
\begin{bmatrix}
\beta_1 \\
\beta_2 \\
\beta_3 \\
\beta_4
\end{bmatrix}
$$

where $D = a_1^2 (1 + a_1^2) \Sigma x_{1e}^2 \Sigma x_{2e}^2 - (a_1 a_2)^2 (\Sigma x_{2e} x_{1e})^2$, the determinant of the product-cross product matrix (Green). First, consider the sign of $D$. An alternative expression for $D$ is

$$
a_1^2 a_2^2 (\Sigma x_{1e}^2 \Sigma x_{2e}^2 - (\Sigma x_{2e} x_{1e})^2) + a_2^2 \Sigma x_{1e}^2 \Sigma x_{2e}^2.
$$

The term in brackets in (15) is positive by the Cauchy-Schwartz inequality. The second term is obviously positive since all components are squared. Hence, $D$ is positive.

Next, consider the first row of (14) excluding $\frac{1}{D}$, which we have already shown to be positive. The first row contains the bias term for $\beta_1$. The first row is
\[(1 + a_1^2) a_2^2 (\Sigma X_{tc}^2) (\Sigma X_{tr} X_{tc}) - (a_1 a_2)^2 (\Sigma X_{1c} X_{2c}) (\Sigma X_{2c} X_{3c}) \] \[\beta_3 + a_1^2 a_2^2 (\Sigma X_{2c}^2) (\Sigma X_{1c} X_{3c}) - a_1 a_2 a_2^2 (\Sigma X_{1c} X_{2c}) (\Sigma X_{2c} X_{3c}) \] \[\beta_4 \ldots \] (16)

If the omitted variable is income and the commodities are superior goods then from equations (1) and (2), both \(\beta_3\) and \(\beta_4\) are positive. If, in addition, the omitted variable is positively correlated with the included variables then the sums of the cross-products are positive.

Continuing the example, if income is omitted and positively correlated with prices, the coefficient of \(\beta_3\) can be rewritten as

\[(a_1 a_2)^2 [(\Sigma X_{tc}^2) (\Sigma X_{1c} X_{2c}) - (\Sigma X_{1c} X_{2c}) (\Sigma X_{2c} X_{3c})] + a_1^2 (\Sigma X_{2c}^2) (\Sigma X_{1c} X_{2c})\] (17)

The second term in (17) is positive by the above assumptions. The term in brackets in (17) is just the numerator of the least squares estimator of \(\alpha_3\) (the coefficient of \(X_3\)) obtained by the auxiliary regression of \(X_1\) on \(X_2\) and \(X_3\). Thus, the term in brackets is positive if \(X_1\) (say the price of a commodity) is positively correlated with \(X_3\) (income variable or another price).

The sign of the coefficient on \(\beta_4\) can be determined as follows. First consider the coefficient for \(\beta_4\), rewritten as

\[a_1 a_2 a_2^2 [(\Sigma X_{tc}^2) (\Sigma X_{1c} X_{3c}) - (\Sigma X_{1c} X_{2c}) (\Sigma X_{2c} X_{3c})]\] (18)


The term in brackets can be interpreted as the least squares coefficient on $X_3$ obtained from the auxiliary regression of $X_1$ on $X_2$ and $X_3$. Thus, the least squares coefficient of $X_3$ will be positive if $X_3$ is another price or income. The only remaining term is $a_1 a_2 a_2^2$. Now $a_1 a_2$ must satisfy the restriction (Kmenta, p. 641)

$$a_1 \sigma_{11} + a_2 \sigma_{12} = 0 \tag{19}$$

where $\sigma_{11}$ is the variance of $\epsilon_{1t}$ and $\sigma_{12}$ is covariance between $\epsilon_1$ and $\epsilon_2$.

Clearly $\sigma_{11}$ is positive. If $\sigma_{12}$ is negative, then (19) implies that $a_1$ and $a_2$ are the same sign. Consequently, the bias term for $\beta_1$, the SUR estimator when $X_3$ and $X_4$ are omitted, must be positive. If $\sigma_{12} > 0$, as would be expected if the two equations were demand functions, then (19) requires that $a_1$ and $a_2$ be of opposite signs. Then, the term involving $\beta_4$ in the bias expression is negative by (18). Thus, the sign of the overall bias term associated with the SUR estimator of $\beta_1$ depends upon the relative magnitudes of the two terms in (16). It is difficult to state a priori the direction the bias would be.

Now, consider the second row of (14). This row contains the bias associated with estimating $\beta_2$. The bias is given by

$$
\frac{1}{D} \left\{ -\left(1 + a_1^2\right)(a_1 a_2) \left( \Sigma X_{1t} X_{2t} \right) \left( \Sigma X_{1t} X_{3t} \right) + a_1 a_2 \left( 1 + a_1^2 \right) \Sigma X_{1t}^2 \left( \Sigma X_{2t} X_{3t} \right) \right\} \beta_3 \\
+ [- (a_1 a_2)^2 \left( \Sigma X_{1t} X_{2t} \right) \left( \Sigma X_{1t} X_{3t} \right) + a_2^2 \left( 1 + a_1^2 \right) \left( \Sigma X_{1t}^2 \right) \left( \Sigma X_{2t} X_{3t} \right) \beta_4] \tag{20}
$$
The coefficient of $\beta_3$, excluding $1_D$, can be rewritten as

$$a_1a_2\left[(\Sigma X_1^2) (\Sigma X_2X_1) - (\Sigma X_1X_2) (\Sigma X_1^2X_2)\right] - a_1a_2(\Sigma X_1X_2) (\Sigma X_1X_2) .$$

(21)

The term in brackets is the numerator of the least squares estimator of $X_2$
obtained by regressing $X_3$ on $X_1$ and $X_2$, e.g., by regressing the omitted income
variable on included prices. The term in brackets will be positive in most
demand estimation contexts. If the other variables are positively correlated,
then the sign of the $\beta_3$ coefficients depends upon the sign of $a_1a_2$. As
before, if $\sigma_{12} > 0$, then $a_1$ and $a_2$ have opposite signs and the coefficient is
positive. On the other hand if $\sigma_{12} < 0$, then $a_1$ and $a_2$ will have the same
signs and the coefficient of $\beta_3$ will be negative given the positive
correlation of the explanatory variables.

Finally, the coefficient of $\beta_4$ can be rewritten as

$$(a_1a_2)^2[(\Sigma X_1^2) (\Sigma X_2X_1) - (\Sigma X_1X_2) (\Sigma X_1^2X_2)\right] + a_2^2(\Sigma X_1^2) (\Sigma X_2X_1) .$$

(22)

By using similar arguments as those above, this term will be positive, again
making the same assumptions concerning the correlation between the explanatory
variables.

The bias associated with the SUR estimator of $\beta_2$ can be, in general,
either positive or negative; however, in the special case developed the sign
can be determined to be positive. In the studies by Capps, and Funk, Meilke
and Huff (or for similar results), since the commodities are superior, one
would expect the income elasticities to be positive. In addition, income is likely to be positively correlated with prices. Thus, omission of income implies that the estimates of cross-price elasticities are biased positively while the own-price elasticity estimate is biased negatively, i.e., the estimated cross-price elasticities are larger when income is omitted than when it is included in the equation.

This result implies that the estimators found in the studies by Capps, and Funk, Meilke and Huff are not consistent with those from more conventional data bases, and that the direct comparisons with results of specifications that include income are not merited. In short, the bias should be investigated prior to the use of the estimated elasticities in pricing and related applied contexts.

Efficiency

For efficiency (and in the context of the model under consideration---equations (1) and (2)), when total "meat" expenditures are omitted from specifications using scanner data, the covariance matrix is

$$\text{COV} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \sigma^2 (X^T X)^{-1}$$

(23)

where

$$X = \begin{pmatrix} X_1 & 0 \\ a_1 X_1 & a_2 X_2 \end{pmatrix}$$
(see the first term in equation (8)). The covariance of $b_1$ (the estimated price elasticities) is the upper left block of equation (23) which, after some algebraic manipulation, becomes

$$\text{COV}(b_1) = \sigma^2 [ (1 + a_1^2) X_1^T M_2 X_1 ]^{-1}$$

where $M_2 = I - (a_1^2/(1 + a_1^2)) X_2 (X_2^T X_2)^{-1} X_2^T$.

If meat expenditures from the stores under investigation are used as a proxy for total meat expenditures as in Capps and Nayga, what would be the variance of $b_1$ compared to that obtained in equation (24)? An analytical solution is beyond the scope of this discussion. Nevertheless, it is well known that the variance of a restricted estimator is always less than or equal to the variance of an unrestricted estimator (Fomby, et al.). Thus, there is a trade off between a biased estimator with smaller variance and an unbiased estimator with a larger variance. Aigner has shown that the size of the mean squared error (MSE) of the remaining coefficient estimates when a proxy is used with possible measurement error can either be larger or smaller than the MSE of the remaining coefficient estimate when the relevant variable is omitted.

If the analyst knows that consumers purchase all or most of their meat from the stores under investigation, then the proper approach, based on the MSE criterion, would appear to be to include total meat expenditures. In this case the measurement error would be small. Alternatively, it is unclear whether the analyst can obtain estimators with more desirable sampling properties by employing a proxy with substantial measurement error. The apparent size of the bias in the two scanner studies, together with the large
sample sizes that are apparently available would, however, suggest serious consideration of using the "available" total meat expenditures variable.

Conclusion

One should conclude that the empirical results reported by Capps, and Funk, Meilke and Huff may be an anomaly. The unusual magnitudes of the estimated own-price elasticities relative to their estimated cross-price elasticity counterparts is likely due to a specification error and not the behavior of consumers. Total expenditures probably should have been included as an explanatory variable as in the Capps and Nayga study when scanner data are employed to avoid specification biases. As shown in the last section, the mean squared error of the estimators of the elasticities can only possibly be improved even assuming that a proxy is used for the true meat expenditures variables; consumers purchase meat elsewhere than in the scanner stores. With these results future users of scanner data in empirical demand analyses should have a stronger basis for interpreting their results relative to estimates of elasticities from alternative sources.
References


