Economics Bulletin

Volume 33, Issue 1

The Outside Good Bias in Logit Models of Demand with Aggregate Data

Dongling Huang
Rensselaer Polytechnic Institute

Christian Rojas
University of Massachusetts Amherst

Abstract

The logit model is the most popular tool for estimating demand in differentiated products markets. However, in its aggregate version, practitioners have to "guess" the size outside good. We propose a way to remove the bias created by an inaccurate guess in simpler versions of the model.

Rensselaer Polytechnic Institute (huangd3@rpi.edu) and University of Massachusetts Amherst (rojas@resecon.umass.edu). Corresponding author: Huang.
Citation: Dongling Huang and Christian Rojas, (2013) "The Outside Good Bias in Logit Models of Demand with Aggregate Data", *Economics Bulletin*, Vol. 33 No. 1 pp. 198-206.
Contact: Dongling Huang - huangd3@rpi.edu, Christian Rojas - rojas@resecon.umass.edu.
Submitted: July 26, 2012. **Published:** January 23, 2013.

1. Introduction

In the logit model the outside good option plays a crucial role because it allows for the possibility that consumers may change their consumption of the whole differentiated good category (Berry, 1994; p.247). A simultaneous price increase caused by a cost shock or a simultaneous decrease in quality illustrate of the importance of specifying an outside good. However, with aggregate data the size of the outside good is unobservable. The customary procedure is to pin down this variable by making an educated guess for the "market potential" (i.e. the number of potential purchases that *could* have occurred in a market). In this paper we illustrate the bias that can be created by an inaccurate market potential guess (both in simulation as well as in an empirical application) and propose a way to ameliorate it in simpler versions of the model.

2. The Bias

Why is there a Bias?

We describe the problem and the proposed remedy using a version of the simple logit that includes one observed exogenous product characteristic (x) and price (p). We assume that in market t, a representative individual i's probability of purchasing brand j is given by $\pi_{jt} = \exp(\delta_{jt})/[1+\sum_{k\neq 0}\exp(\delta_{kt})]$, $(j,k=0,1,\ldots,J)$; where $\delta_{jt} = \beta_0 + \beta_1 x_{jt} + \alpha p_{jt} + \varepsilon_{jt}$ is the mean utility and ε_{jt} is an unobserved product characteristic correlated with price. We follow the practice of studies with aggregate data and: a) normalize the utility of the outside good to zero ($\delta_{0t}=0$), and b) use observed market shares, $s_{jt} = q_{jt}/\sum_{k=0}^{J} q_{kt}$, in lieu of π_{jt}. As shown by Berry, instrumental variable treatment of ε_{jt} is possible through the inversion of mean utilities:

$$ln(s_{jt}) - ln(s_{0t}) = \delta_{jt} \qquad (1)$$

In the absence of data for the outside good (q_{0t}), practitioners carry out estimation of (1) by using a 'reasonable guess' for the true market potential, $M_t = \sum_{k=0}^{J} q_{kt}$. Note that this is equivalent to formulating a guess for q_{0t}. In what follows, we use the tilde notation to differentiate the true values (M_t and q_{0t}) from their "guessed" counterparts ($\tilde{M}_t$ and $\tilde{q}_{0t}$). In other words, the outside good enters (1) with error: $v_t = \ln(q_{0t}) - \ln(\tilde{q}_{0t})$. Equation (1) can then be rewritten as:

$$ln(q_{jt}) - ln(\tilde{q}_{0t}) = \delta_{jt} + v_t \qquad (2)$$

It is easy to see that the term v_t is a function of the outside good and thus correlated with the explanatory variables in δ_{jt} since $s_{0t} = 1/[1+\sum_k \exp(\beta_0 + \beta_1 x_{kt} + \alpha\ p_{kt} + \varepsilon_{kt})]$.[1] Note that an identical argument holds for the more flexible nested logit model since equation (2) only needs a

[1] Using a Taylor series expansion it can be shown that the correlation between v_t and x_{jt} is equal to $\sum_{l=1}^{\infty}\left[Cov\left(\left(\sum_{k=1}^{J} q_k\right)^l, x_j\right)\cdot\left(\frac{1}{\tilde{M}^l} - \frac{1}{M^l}\right)\right]/l$. Therefore, the magnitude of the bias can be thought of (or analyzed) as in an omitted variables problem. Proof available upon request.

slight modification in the mean utility: $\delta_{jt} = \beta_0 + \beta_1 x_{jt} + \alpha p_{jt} + \sigma \ln(s_{j/g}) + \varepsilon_{jt}$, where $\ln(s_{j/g})$ is the market share of each product j with respect to all products in nest g and σ measures the correlation of utility across products in g.

Importantly, usual IV techniques used to address the endogeneity of price (or other explanatory variables) will not alleviate the potential bias caused by v_t. Specifically, since instruments are chosen so that they are correlated with price and since v_t is also a function of price, such instruments will fail to be orthogonal to v_t. This is clearly illustrated in the simulations below, where instrumentation of price does not eliminate the bias.

Should We Worry About the Bias?

Since it is not possible to derive an analytical expression for the bias caused by v_t, we illustrate the possible consequences via simulation.[2] We consider 1,000 five-product markets (i.e. J=5 and T=1,000) with $\beta_0 = 0.5$, $\beta_1 = 0.5$ and $\alpha = -1.5$. Endogeneity is introduced by setting $p_{jt} = \tilde{p}_{jt} + z_{1jt} + z_{2jt}$ and $\tilde{\varepsilon}_{jt} + z_{2jt}$, where $\tilde{p}_{jt} \sim U(1, A)$, $\tilde{\varepsilon}_{jt} \sim N(0,0.25)$, $z_{1jt} \sim N(0,0.09)$ and $z_{2jt} \sim N(0,0.09)$.[3] The variable x_{jt} is assumed to be exogenous and dichotomous (e.g. a product feature): $x_{jt} \sim I[U(0,1) > B]$, where $I[\cdot]$ is equal to one when its argument is true. We set different A and B parameters for each of the five goods to allow for an asymmetric market: A=(2.5, 1.5, 2, 3, 3.5) and B=(0.7, 0.3, 0.4, 0.5, 0.6).

A simulated dataset consists of 1,000 observations (T=1,000) for 5 inside goods, as well as the outside good. The simulated data take the form of probabilities, which are then transformed to quantities by multiplying the shares by a constant M (for illustration purposes we choose M=100 for all t). We simulate 100 datasets with this parameterization and exclude q_{0t} prior to estimation. Finally, for each simulated dataset, we carry out standard instrumental variable estimation of (1) with 400 different assumed market potentials, some smaller and some bigger than the true M. We denote the a^{th} assumed market potential as $\widetilde{M}_a$.[4]

We report four figures from this simulation. All figures report a growing $\widetilde{M}_a$ on the x-axis. Figure 1 plots the correlation between (x_{jt}, p_{jt}) and v_t (median across the 100 datasets). Figure 2 reports the estimated price coefficient. Figures 3 and 4 report the estimated own-price elasticity for product 1 and cross-price elasticity between products 1 and 2, respectively.

It can be seen that the correlation between (x_{jt}, p_{jt}) and v_t is always statistically different than zero and that it depends on the size of the assumed market potential. Further, as expected, this correlation depends on the sign of the coefficient of the independent variable. Also, the sign of the correlation switches at the true market potential because v_t also changes in sign at this threshold.

Two patterns in figures 2 through 4 were consistently noted in other variants of this Monte Carlo exercise. First, the bias in the estimated price coefficient (α) and the own-price elasticity tends to be moderate. Second, the bias can be large for the cross-price elasticity.

[2] The bias in the reported simulation is commonly observed in other variants the exercise. For example, the simulations in Huang, Rojas and Bass (2008) illustrate the bias in a setting similar to that of Berry.

[3] We choose the variances of z_{1jt}, and z_{2jt} so that their contribution to the variance of $\tilde{p}_{jt}$ is not too large.

[4] The smallest market potential is defined by $\widetilde{M}_1 = \max_t(\sum_j q_{jt})$. Each subsequent $\widetilde{M}_a$ grows by 0.5.

Figure 1: Correlation between v_t and independent variables for different $\widetilde{M}_a$, with 95% confidence intervals

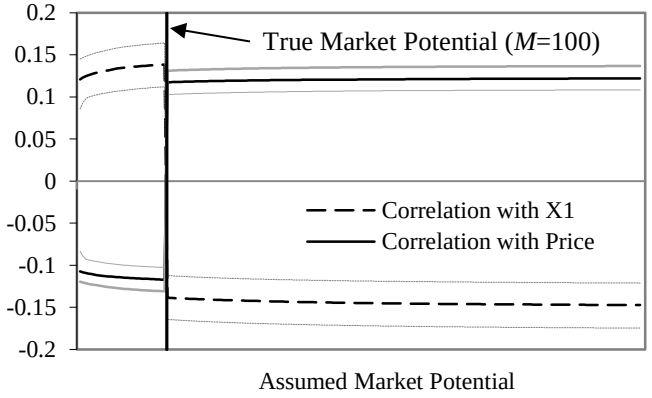
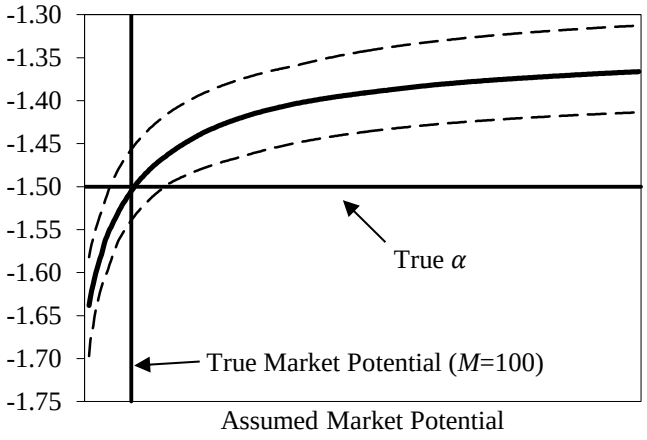


Figure 2: Estimated α for different $\widetilde{M}_a$, with 95% confidence intervals



Note: standard IV estimation is applied throughout (i.e. figures 2, 3 and 4).

Figure 3: Estimated Own-Price Elasticity for different $\widetilde{M}_a$, with 95% confidence intervals

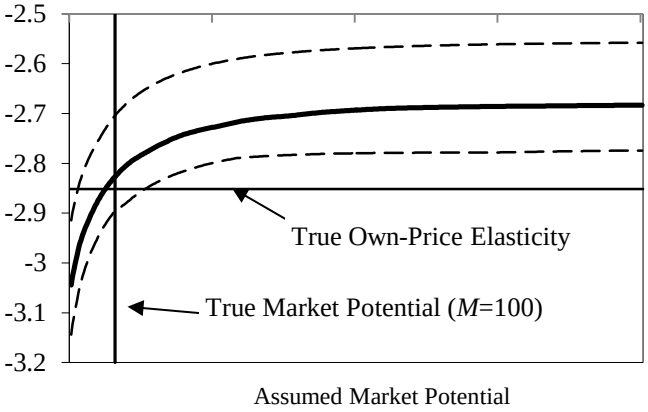
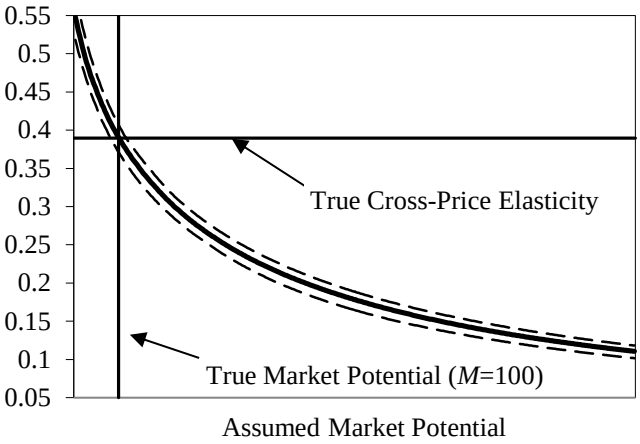


Figure 4: Estimated Cross-Price Elasticity for different $\widetilde{M}_a$, with 95% confidence intervals



Practitioners could be tempted to conclude that the bias might not be worrisome if the price coefficient is not too sensitive to $\widetilde{M}$. However, the useful measure for most empirical work is price elasticity, precisely where the bias could be severe. A closer look at logit elasticity formulas (shown below) explains the noted patterns:

$$\rho_{jk} = \begin{cases} \alpha p_j (1 - s_j) & if\ j = k \\ -\alpha p_k s_k & if\ j \neq k \end{cases}$$

Both own- and cross-price elasticities depend on the assumed market potential through two channels: the estimated α as well as market shares (s_j and s_k).[5] As it can be seen in figure 2,

[5] Strictly speaking s_j (and s_k) in the elasticity formulas should correspond to predicted shares (i.e. $\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{jt} + \hat{\alpha} p_{jt}) \,/\, [1 + \sum_{k \neq 0} \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{kt} + \hat{\alpha} p_{kt})]$, and not the 'observed' shares (i.e. $q_j \,/ \widetilde{M}$). In practice, however, predicted and observed shares are related to $\widetilde{M}$ in similar ways. The reason for this is that $\widetilde{M}$ is inversely related to

when $\widetilde{M}$ is large, the estimated α is biased towards zero (and vice versa). This means that a large (small) $\widetilde{M}$ biases both elasticities towards (away from) zero. However, the effect of an estimated α that is biased towards (away from) zero is counteracted by a corresponding smaller (larger) s_j in the case of the own-price elasticity but it is magnified in the cross-price elasticity by a corresponding smaller (larger) s_k.

A similar situation arises with nested logit elasticities, which we display below. Specifically, for a given σ, the bias introduced by α is magnified in both cross-price elasticity formulas (second and third lines in the formula) by nest shares s_g, which are an inverse function of $\widetilde{M}$. The opposite occurs for own-price elasticity.

$$\rho_{jk} = \begin{cases} \dfrac{\alpha}{(1-\sigma)} p_j\left(1 - \sigma s_{j/g} - (1-\sigma) s_{j/g} s_g\right) & if\ j = k \\ -\dfrac{\alpha}{(1-\sigma)} p_k s_{k/g}\left(\sigma + (1-\sigma) s_g\right) & if\ j \neq k\ and\ j,k \in g \\ -\alpha p_k s_{k/g}, s_{g'} & if\ j \neq k\ and\ j \in g, k \in g' \end{cases}$$

To be clear, as with any simulation exercise, the patterns of the bias observed in the following simulation should not be interpreted as a general rule. In fact, our empirical application below shows that the bias could be severe for both own- as well cross-price elasticities.

3. The Proposed Solution

Our solution allows the researcher to recover an unbiased estimate of α by treating the unobserved outside good quantity, q_{0t}, as a market fixed effect. Specifically, one can write (1) as $ln(q_{jt}) = \delta_{jt} + ln(q_{0t})$ and difference out the unobservable:

$$ln(q_{jt}) - ln(q_{kt}) = \delta_{jt} - \delta_{kt} = \beta_1(x_{jt} - x_{kt}) + \alpha(p_{jt} - p_{kt}) + (\varepsilon_{jt} - \varepsilon_{kt}) \qquad (3)$$

In other words, within each market, one can arbitrarily choose any *k* as the "differencing" product and carry out estimation of (3) via standard instrumental variable methods; we call the estimated vector $\hat{\theta}_{IV}^{d}$, where the superscript *d* denotes the "differenced" nature of the data. A similar situation arises in the nested logit case where the right-hand side of (3) contains the additional term $\sigma\left(\ln\left(s_{j/g}\right) - \ln\left(s_{k/g}\right)\right)$. To illustrate, using (3) in our simulation above yields and estimate of α equal to -1.50 (s.e.=0.03).

Our solution, however, is only partial because the parameter β_0, which is needed to compute purchase probabilities (or market shares in the aggregate version), is unidentified. This is problematic because elasticities are a function of market shares. To complete the proposed fix, we suggest a mechanism, based on the monotonic bias we observe, to obtain a ballpark estimate of market shares. Specifically, the researcher can first carry out estimation of (3) to back out the unbiased estimate of $\hat{\theta}_{IV}^{d}$ and then search for the $\widetilde{M}$ that yields estimates of α and β_1 that are closest to the unbiased value.

$\hat{\beta}_0$ (not shown here): a large $\widetilde{M}$ results in a small $\hat{\beta}_0$ and consequently in small predicted shares. Elasticities reported here are computed with predicted shares.

Specifically, we propose the following minimum distance estimator:

$$\widetilde{M} = \underset{M}{\operatorname{argmin}} (\hat{\theta}_{IV}^{l}(M) - \hat{\theta}_{IV}^{d})' (\hat{\theta}_{IV}^{l}(M) - \hat{\theta}_{IV}^{d}) \quad (4)$$

where $\hat{\theta}_{IV}^{l}(M) = (X'P_Z X)^{-1}(X'P_Z Y)$, $P_Z = Z(Z'Z)^{-1}Z'$, Y is a vector with element $y_{jt} = \ln(\frac{q_{jt}}{(M - \sum_1^J (q_{jt}))})$, X is a matrix of explanatory variables, and Z is the usual matrix of instruments.[6] That is, this estimator searches for the value of M that produces the estimates that are closest to $\hat{\theta}_{IV}^{d}$. For statistical inference of this estimate, we rely on a bootstrap technique.[7] Applying this estimator to our simulation above, we obtain a median of $\widetilde{M} = 96.08$ with 95% confidence intervals of [89.13, 105.54].

The proposed remedy correctly identifies the unique correct market potential in our simulations (see figures 2 through 4). Here, identification is achieved through a feature of the data, as well as by functional form. First, notice that the outside good will become more attractive when the inside goods (as a whole category) are less attractive (for example the average price of inside goods is higher), and vice versa. In the data, the variation of prices and other characteristics of the inside products allows us to capture the relative attractiveness of the whole product category with respect to the outside good; this variation, although important, is only useful for capturing the substitution towards the no purchase category and thus, alone, cannot determine the size of market potential. Second, the functional form allows us to treat the market potential (and thus the size of the outside good) as a market fixed effect effectively eliminating the bias (in α and β_1) that an incorrect guess might cause. A second (and more critical) feature given by functional form allows us to pin down the *level* of *M*: the monotonic relationship between the assumed market size and parameter estimates.

The proposed solution, however, has some limitations. In our simulation, *M* is fixed across markets (by design), but in practice, this quantity could move from one market to the next. For example, consumers in one city may have a higher tendency to consider other product categories; alternatively, the product category under study has become increasingly less popular over time. Because of this, our proposed solution for the choice of $\widetilde{M}$ does not constitute a solution that can be readily applied to all empirical problems.

Our approach can be relaxed to accommodate for a more "flexible" search of $\widetilde{M}$ thereby making the solution implementable in many applications. Specifically, one can specify *M* as a proportion of a variable that is likely to determine market potential across different time periods or over regions. For example, one could define market potential as a proportion of the total population in the market (e.g. Berry, Levinsohn and Pakes, 1995; Nevo, 2001) and carry out the search in the second step of our suggested method by changing such proportion. The attractiveness of this option, besides being more in line with what one would expect in real markets, is that market potential varies over markets in an economically meaningful way.

[6] Since the coefficient on constant is not identified in (3), and therefore not contained in $\hat{\theta}_{IV}^{d}$, we exclude the coefficient on the constant in $\hat{\theta}_{IV}^{M}$ to carry out the estimation in (4).

[7] Specifically, we draw as many markets as we observe in the data (with replacement) and first estimate $\hat{\theta}_{IV}^{d}$ using equation (3). Then, we estimate $\widetilde{M}$ using equation (4). We perform this exercise multiple times and define confidence intervals in the customary fashion (Wooldridge 2009, p. 223). We have worked on obtaining an analytical solution for the asymptotic distribution of $\widetilde{M}$, but we are uncertain about its existence. If feasible, we expect to incorporate the analytical solution in future work.

Table I illustrates results from this procedure using Nevo's data. We use Nevo's specification 9 from Table V in his paper and carry out regressions using different number of servings per week (i.e. the proportion), including 7 servings as assumed in his study.[8] The last column of the table contains results of applying (3), while the second specification results from applying the $\widetilde{M}$ found when applying equation (4). For the sake of brevity, we report the estimates of the price and advertising coefficients, together with the corresponding median own- and cross-price elasticities.

Consistent with our simulations, we observe a monotonic relationship between $\widetilde{M}_a$ and the estimated coefficients and elasticities. For estimation and inference of $\widetilde{M}$, we apply (respectively) equation (4) and the bootstrap procedure described in footnote 7. This exercise suggests that the appropriate size of the market potential should be 2.48 servings (95% confidence intervals of [2.17, 2.95]) with resulting median own- and cross-price elasticities that are larger (in absolute value) by 81.17% and 420.8%, respectively, than those obtained with 7 servings/week.

Table I: Logit Results in Nevo's Study, Different Market Potentials and Proposed Regression, (s.e.)

	Spec. 1	*Spec. 2*	*Spec. 3*	*Spec. 4*	*Spec. 5*	*Spec. 6*	*Spec. 7*
Price	-59.530	-38.096	-25.043	-20.799	-19.566	-19.137	-36.905
	(1.722)	(1.094)	(0.792)	(0.703)	(0.679)	(0.670)	(0.720)
Advertising	0.005	0.017	0.024	0.026	0.026	0.027	0.017
	(0.005)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Median own-price elasticity	-11.21	-7.21	-4.77	-3.97	-3.75	-3.67	N/A
Median cross-price elasticity	0.242	0.125	0.051	0.024	0.016	0.013	N/A
# Servings/week	2	2.53	4	7	10	12	N/A
Estimated Equation	(1)	(1)	(1)	(1)	(1)	(1)	(3)

4. Discussion

Our results using the simple logit version suggest that large elasticity biases (both own and cross) can arise as a result of an inadequate assumed market potential. This can have severe consequences for inference. For example, if the assumed market potential is too large in a merger simulation, the artificially low cross-price elasticities could lead to an underestimation of the post-merger price increase.

Our proposed remedy is similar to much of the current empirical practice in which the researcher checks the sensitivity of results to the size of assumed market potential. The difference is that in our approach the researcher knows what he/she is looking for. The solution we propose cannot be applied in the BLP algorithm that is used for the random coefficients case because the measurement error in market potential enters the algorithm in a non-linear fashion. Although our focus is on the simple logit and nested logit versions, we believe that the proposed remedy can be used to generate a market potential guess in the random coefficients variant that is

[8] Our results are not identical to Nevo's, but they are quite close.

more educated than what is dictated by the current practice.[9] Specifically, practitioners can use the suggested approach in the simpler versions to obtain a ballpark estimate of the market potential; this guess can then be used in the full-fledged model. Our current and future work is aimed at extending the proposed solution to the more general model.

[9] We are currently investigating another approach, suggested by a referee, that relies on indirect inference arguments (see Gourieroux, Monfort and Renault, 1993).

5. References

Berry, S. (1994) "Estimating Discrete Choice Models of Product Differentiation" *RAND Journal of Economics* **25**, 242-62.

Berry, S., Levinsohn and A. Pakes (1995) "Automobile Price in Market Equilibrium" *Econometrica* **63**, 841-89.

Gourieroux, C., A. Monfort and E. Renault (1993) "Indirect Inference" *Journal of Applied Econometrics* **8**, 85-118.

Huang, D., C. Rojas and F. Bass (2008) "What Happens When Demand is Estimated with a Misspecified Model?" *Journal of Industrial Economics* **56**, 809-39.

Nevo, A. (2001) "Measuring Market Power in the Ready-to-Eat Cereal Industry" *Econometrica* **69**, 307-42.

Wooldridge, J. (2009) *Introductory Econometrics: A Modern Approach*, 4th Edition, Cengage Learning.