A Hurdle Model Estimation of Household and Product Characteristics to Explain Online Buying Decisions

MEYYAPPAN NARAYANAN, BONWOO KOO, AND BRIAN P. COZZARIN

March 2008

Working Paper

Key Words: online buying, search vs. experience goods, two-step estimation, hurdle model, complementary log-log model, truncated regression, truncated Poisson regression.

JEL codes: D10, L16, L81.

Meyyappan Narayanan is PhD student, Bonwoo Koo is Assistant Professor, and Brian P. Cozzarin is Associate Professor, all in the Department of Management Sciences, University of Waterloo, Canada. They may be contacted at mnarayan@uwaterloo.ca, bonkoo@uwaterloo.ca, and bpcozzar@uwaterloo.ca, respectively.

The authors thank Mary E. Thompson for her extensive help; the study may not have come to fruition without it. The authors also thank Scott A. Jeffrey for his numerous comments on an earlier draft of the paper.

Thanks are also due Ms. Carol Perry and Ms. A. Michelle Edwards, Ph.D., of University of Guelph for their help with the datasets of the Tri-University Data Resources (TDR).

Abstract

This study employs a hurdle model procedure to estimate demographic and product category variables to explain household online buying decisions. Households decide in the step-one whether to buy online and if decided to buy online, further decide in the step-two how intensely to buy. These two decisions may be influenced by different variables or to different extent by the same variables, hence the two-step procedure. We employ Probit and Complementary Log-Log models for the step-one, and Truncated Regression and Truncated Poisson models for the step-two on a large dataset of the Household Internet Use Surveys of 1999 – 2003, compiled by Statistics Canada. The study finds that the online buying probability is the highest for households with young head of household, but buying intensity is the highest for households with head of household in the 55 – 64 age group, possibly because such households are large or have teen-agers driving buying intensity. It further finds that experience goods are bought more heavily than search goods and that dollar value per purchase is the highest for experience goods.

1. Introduction

It is ever more important to understand household online buying behavior. Online buying can be beneficial (e.g., Bakos, 1997; Borenstein and Saloner, 2001; Vulkan, 2003), so more and more households are adopting online buying (e.g., Bakos, 2001; Michalak and Jones, 2003). Almost all households – except perhaps those in remote areas - seem to have computer and to be connected to the Internet, and the necessary setting for online buying to spread and grow among households seems to be firmly in place at least in the developed countries.

Indeed there has been quite an amount of research undertaken by researchers to understand the predictors of online buying adoption (e.g., Bellman, et al. 1999; Lohse, et al. 2000; Goolsbee, 2000; Wolfinbarger and Gilly, 2001), buying intensity (Lohse, et al. 2000; Miyazaki and Fernandez, 2001; Li, et al. 2003), and buying intention (Heijden, et al. 2001; Park, 2002). These studies have used non-demographic consumer variables such as "looking for product information," "wired lifestyle," "the amount of discretionary time households have," and shopping orientations; demographic variables as main variables or as controls (e.g., Hoffman, D. L. and Novak, T. P. 1999); seller variables such as website design, privacy concerns, and trust; and market variables such as local sales tax rates and product types.

Though these studies have uncovered various insights about online buying behavior, they do not give adequate insights about the relation between household demographics and online buying decisions. Some have concluded that demographics are unimportant pointers to online buying decisions ("Look for a wired lifestyle and time starvation, not demographics" - Bellman, et al. 1999). Other studies are simply contradictory: probability of buying is high when income, education, and age are high (Bellman, et al. 1999); high when income and education are high, but when age is low (Goolsbee, 2000); significantly explained only by gender (Lohse, et al. 2000); buying frequency is high when income is high (Bellman, et al. 1999); value of purchases is significantly explained only by income (Lohse, et al. 2000); only by education (Li, et al. 2003).

Some possible reasons for the above contradictions concern estimation methods used. While the studies rightly used logistic regression to model probability of buying online and classic regression (e.g. Ordinary Least Squares) to model log value of purchase (which was shown in Bellman et al. (1999) to approximately follow normal distribution), Bellman et al. (1999) seem to have used similar regressions to model both log value of purchase and number of transactions (buying frequency), which is another measure of buying intensity[1].Frequency is a count-type variable for which Poisson regression is one of the appropriate methods. When observations with zero value for the dependent variable are excluded from regression, Truncated Poisson regression is appropriate. Another possible method-related problem is correlation among explanatory variables. Ideally explanatory variables should be identified from theory; they should be exogenous in any case. When the explanatory variables are more "fundamental," they are unlikely to be endogenous and correlated among themselves. Correlation among explanatory variables will result in wrong estimates. Other explanatory variables used in Bellman, et al. (1999) and Lohse, et al. (2000) may be correlated with demographic variables. For example, "wired lifestyle" may be correlated with age since young people may be spending more time on the Internet; similarly "searched the Internet for finance" and "time-starved" may be correlated with income (since high-income people may be more likely to search finance sites or to be busier at their career). It may be noted that the demographic variables are likely to be more "fundamental" than the above other variables since they are likely to be drivers, not driven.

Even if demographic variables are not the primary predictors of probability and intensity of online buying, a proper study of demographic characteristics of households is essential in order to more completely understand household online buying decisions since there is reason to believe that

[1] The details of the regression used in Li, et al. (2003) are not stated.

peoples' tastes relate to their demographic characteristics in addition to any other factors. Hence this study undertakes estimation of demographic characteristics to explain online buying decisions by households with respect to adoption and buying intensity, using two separate measures for buying intensity, namely, log value of orders and number of orders for online purchase by households. In a nice study in the context of agricultural marketing contract decisions by farms, Katchova and Miranda (2004) employed two-step estimation methods involving probit and truncated regressions for the steps one and two in the model with continuous-type dependent variable, and complementary log-log and truncated Poisson models for the two steps in the model with count-type dependent variable. Following them, we employ this two-step hurdle model approach to our estimations.

Though studies of online buying decisions and intention have focused on consumer variables or seller variables (e.g. website design) studies have also considered product characteristics as predictors. However, many such studies really considered product groups (such as music, books, hotel reservation), not product categories, except some such as Asch D. (2001); Bei, L. et al. (2004); and Korgaonkar, P. et al. (2006). Product categorization propounded by Nelson P. (1970, 1974) and Darby & Karni (1973) are widely accepted and followed (Kline, L.R. 1998). We include product category variables accordingly in our estimations so we can get further insights about online buying decisions.

The paper is henceforth organized as follows: section 2 presents a theoretical model of household online buying; section 3 discusses the hurdle model used for the estimations; we then explain our data, and present and discuss the estimation results in section 4; and section 5 concludes. The descriptive statistics, estimation results and our categorization of product groups are appended.

2. Theoretical Model

We model the choice between online buying and buying at local store as follows[2].

Assumptions

- A consumer (household) will buy one unit of a good.
- She has a choice of buying it either online or in the local store.
- The good's quality parameter is defined as *s*.
- The individual consumer's taste parameter is q, which varies by consumers.
- The distribution of q in the population is F(q).
- The effort parameter for information search is a.
- The price of the good is p.

Model

If the consumer purchases the good online, her utility is

$$U_e = q\, s_e - p_e - a_e \qquad (1)$$

If she purchases the good in the local store, her utility is

$$U_n = q\, s_n - p_n - a_n \qquad (2)$$

[2] The model is adapted from Tirole (1989).

The consumer will buy online if

$U_e > U_n$, or

$$U_e - U_n = q\,(s_e - s_n) - (p_e - p_n) - (a_e - a_n) > 0 \quad (3)$$

We can redefine the parameters as the difference between online buying and store buying:

$$U = q\,s - p - a \quad (4)$$

Assume there are N consumers (households) in the population. Consumers whose utility is greater than 0 or

$$q > (p + a)/s \quad (5)$$

will demand the good online. Thus the demand function is

$$D(p) = N[1 - \mathrm{F}((p + a)/s)] \quad (6)$$

Interpretation

Quality parameter *s*:
- For search goods[3], $s_e \leq s_n$ and $s \leq 0$. Buying at local store permits more complete searching because physical inspection is possible. For example, a shirt can be tried at local store.
- For experience goods, $s_e \geq s_n$ and $s \geq 0$. Buying at local store does not have any particular advantage over buying online because an experience good's quality can not be ascertained before purchase even with physical inspection. On the other hand, online buying may be advantageous because of more choices for the consumer.

Price of the good *p*:
- Price (p_e or p_n) includes transportation cost and sales tax (which is 0 for online buying if it is evaded).
- Online price can be more (e.g., price discrimination) or less (i.e., at a competitive level due to large number of suppliers).
- On the other hand, the price at the store is at the level of local monopoly.

Effort parameter *a*:
- This is cost for information collection. An example of low cost for information collection is searching the Internet through a high-speed connection, so the consumer spends less time searching.

Analysis

The inequality (5) models the online buying decision. Any consumer whose taste *q* is greater than the threshold $(p + a)/s$ will buy the good online. When *s* is 0, $(p + a)$ must be negative in order to

[3] Nelson (1970) classified goods as search goods (those whose quality can be fully ascertained before purchase – e.g., shirts) and experience goods (those whose quality can be ascertained only after purchase, that is during consumption – e. g., hotel reservations). Darby & Karni (1973) later added the category credence goods (those whose quality can not be ascertained even after consumption – e.g., surgeries)

induce online buying, as can be seen from the equation (4). On the right-hand side of (5), any price advantage and effort cost advantage will decrease the numerator of the threshold; any quality advantage, as may be in the case of an experience good, will increase the denominator. It is easy to see that online buying is advantageous for goods like hotel reservations. On the left-hand side, the individual's taste parameter may be explained by her demographic characteristics, apart from any psychological or other factors that shape preferences. On the right-hand side, the price advantage (or price sensitivity) may be influenced by consumer's income; the effort cost advantage (or the relative ease with which consumer is able to buy online instead of at local store) may be explained by demographic characteristics (such as age, gender, and education), her shopping orientation, variables like "wired lifestyle," and the type of internet connection; and the quality advantage (or the perception of quality) may be influenced by product as well as demographic characteristics. The inequality can pertain to online buying adoption (first-time buying) or buying intensity (subsequent ongoing purchases). The left-hand side, the right-hand side, and their inter-relationship will vary for each situation. Especially, an individual's effort cost advantage may be considerably different for adoption and subsequent purchases. In other words, the adoption decision and intensity decision may be explained by different explanatory variables or to different extent by the same variables. Moreover, the above inter-relationship may vary over time, according to trends, and depending on states of technologies. Also, we do not clearly know the possible distributions of the taste parameter q. In short, what determines the above inter-relationship is not well understood, as evidenced by different variables advocated by different studies as predictors of online buying behavior. To attempt to understand the phenomenon any further theoretically is beyond the scope of this paper, so we proceed with our empirical tests using demographic and product category explanatory variables as justified in the introduction.

3. Econometric Models

We employ a two-step hurdle model estimation procedure that operates as follows: In the step one, households that are connected to the Internet decide to adopt online buying or not to adopt; if decided to adopt, they decide in the step two how intensely to buy online (that is how much to buy or how frequently to buy in a given period). Log value of orders that households ordered in one year is a measure for how much to buy in one year; and number of orders in one year is a measure for how frequently to buy in one year. We do two different sets of estimations with these two different dependent variables. These two variables are correlated and measure the same thing (online buying intensity) in some sense; however, subtle differences exist in the interpretation of the estimates of the two models, which we need to take care when we analyze the results. Henceforth we discuss the two models separately.

Log Value of Orders Model

Value of orders and log value of orders are continuous variables with range zero to + infinity and – infinity to + infinity, respectively. The log value of orders (like log value of income, for example) has been found to follow normal distribution. In the case of connected households, there is a large proportion that does not buy online. So there are a large number of observations in the sample with zero value for value of orders. In such cases, if we were to use a one-step estimation procedure, we need to limit the dependent variable to non-zero values so we get right estimates. The non-zero values pertain to households that actually bought online. In this case, Tobit regression is an appropriate limited dependent variable estimation method to use.

The structure of the Tobit model is as follows:

$Y_t^* = X'_t ß + e_t$
$Y_t = 0 \quad$ if $Y_t^* \leq 0$
$Y_t = Y_t^* \quad$ if $Y_t^* > 0$

Where Y^* is a latent variable generated by the model, Y is the dependent variable, X is the vector of explanatory variables, $ß$ is the vector of coefficients, and e is the error term assumed to be independently and normally distributed with mean 0 and variance s^2. The index t refers to observation t of the sample. A limit value other than zero can be specified.

The Tobit model has two steps in built. The step one is a Probit model of yes or no adoption decision. The step two is a truncated regression model for the observations kept in the model after censoring[4]. In our case, observations with zero for the value of order are censored. This step two pertains to the continuous decision of how much to buy online. The Tobit model implicitly assumes that the two steps are explained by the same set of explanatory variables and that the estimates are exactly the same in the two steps for each of the variables. This is a restrictive assumption since in reality the two decision processes may be fundamentally different.

The two-step estimation procedure followed by Katchova & Miranda (2004) relaxes the restrictive assumption and allows for the fact that the two decision processes may be influenced by different sets of explanatory variables or to different extent by the same set of explanatory variables. In step one, a Probit model is estimated with a set of explanatory variables. In step two, explanatory variables are changed or kept the same as justified by theory and a Truncated Regression is run on the uncensored observations. It may be noted that if the decision processes of the two steps are indeed different but are explained by the same set of explanatory variables, the Truncated Regression will return point estimates different from the Probit model. It may further be noted that the dependent variable, which is assumed to follow normal distribution in its full range, follows a truncated normal distribution in its limited range in the step two, hence the name Truncated Regression.

We do Probit model estimation for step one and Truncated Regression for step two as above.

There is a Likelihood Ratio (LR) test, as follows, to check if the one-step Tobit estimation is to be rejected in preference to the above two-step estimation:

$LR = 2 * (\ln L_{probit} + \ln L_{truncated\ regression} - \ln L_{tobit})$

where the likelihood ratio statistic LR follows c^2 distribution with r degrees of freedom, r being the number of explanatory variables including a constant and ln L is the log-likelihood value of the respective regression.

Number of Orders Model

Number of orders a household orders in a year is a count-type variable. For a household not adopting online buying, it is zero. The number of orders ranges from 0 to + infinity in integers. Count-type variables may reasonably be assumed to follow Poisson distribution. If we were to use a one-step estimation procedure, Poisson regression may be an appropriate limited dependent variable estimation method to use for number of orders.

[4] For a detailed mathematical exposition of the one-step and two-step procedures, please refer to Katchova & Miranda (2004).

The Poisson regression model has the following structure:

$$\Pr(Y_t = y) = \frac{e^{-\lambda_t}\lambda_t^{\,y}}{y!} \text{ for } y = 0,1,2,\ldots \text{and}$$

$$\lambda_t = e^{X_t'\beta}$$

where $\Pr(Y_t = y)$ is the probability that the dependent variable Y_t (number of orders) equals y (the observed count for number of orders), λ_t is the parameter of the Poisson distribution computed for the observation t (assumed to be the exponent of the linear predictor $X'_t\beta$), X is the vector of explanatory variables, and β is the vector of coefficients. The index t refers to observation t of the sample.

The Poisson regression model composes of two steps: step one pertaining to the Binary Probability model of yes ($y > 0$) or no ($y = 0$) for the online buying adoption decision; step two pertaining to the Truncated Poisson model for all observations kept in the model after truncation (that is, the households that buy online, so $y > 0$). The step two concerns the continuous decision of how frequently to buy online. The Poisson model, like the Tobit model, implicitly makes the same restrictive assumption that the two steps are explained by the same set of explanatory variables and that the estimates are exactly the same in the two steps for each of the variables.

The two-step procedure relaxes the restrictive assumption. The step one model is as follows:

$$\Pr(Y_t = 0) = e^{-\lambda_t}$$

for the probability that a household does not adopt online buying and

$$\Pr(Y_t > 0) = 1 - e^{-\lambda_t}$$

for the probability that a household adopts online buying. Since the Poisson parameter is assumed to be the exponent of the linear predictor, the above model is a complementary log-log model. In step two, explanatory variables are changed or kept the same as done in the two step procedure for the Log Value of Orders model and a Truncated Poisson regression is run on the observations that remain after truncation (that is, the observations with non-zero count for number of orders). The step two model is as follows:

$$\Pr(Y_t = y) = \frac{e^{-\lambda_t}\lambda_t^{\,y}}{y!(1-e^{-\lambda_t})} \text{ for } y = 1,2,\ldots \text{and}$$

$$\lambda_t = e^{Z_t'\gamma}$$

where Z is the vector of (possibly) different set of explanatory variables and γ is the vector of coefficients, which may be the same as or different from β when Z is the same as X and is different from β when Z is different from X. It may further be noted that the dependent variable, which is assumed to follow Poisson distribution in its full range, follows a Truncated Poisson distribution in its limited range in the step two, hence the name Truncated Poisson regression. Here the truncated Poisson probability is conditional probability that $y > 0$.

We do Complementary log-log estimation for step one and Truncated Poisson regression for step two as above for our Number of Orders model.

Again we can use a LR test, as follows, to check if the one-step Poisson estimation is to be rejected in preference to the above two-step estimation:

$LR = 2 * (\ln L_{cloglog} + \ln L_{\text{truncated Poisson regression}} - \ln L_{Poisson})$

4. Data and Estimation Results

HIUS Data

We do estimations using detailed data on the Internet activities of Canadian households, collected by the Science, Innovation, and Electronic Information Division of Statistics Canada for 1999 – 2003[5] through annual surveys known as the Household Internet Use Survey (HIUS)[6]. This survey reports on Canadians using the Internet and measures the extent of their use, location of use, frequency of use, and their reasons for using or not using the Internet. The HIUS has been conducted from 1997 and has evolved to capture increasingly more detail. In 1999, data on electronic commerce (e-commerce) from home were provided. The 2003 survey examined Canadian households' access to the Internet at home, in the work-place, and other locations such as public library, school / university, and Internet café. The collected data reveal relationships between usage and household income, location of use, and demographic factors such as age and education. The detailed set of questions dealing with household e-commerce that were introduced in 1999 was repeated each year thereafter until 2003.

The objectives of the HIUS survey are to, among others, gain a better understanding of how Canadian households use the Internet, identify the types of Internet services used at home, find reasons for non-usage of the Internet, determine what factors would induce households to start using the Internet, understand the impact of the Internet on purchases of goods and services, etc. In assessing the use of the Internet, Statistics Canada has measured the accessibility of the Internet from different locations as well as the frequency and intensity of use from home.

The HIUS survey datasets published by Statistics Canada contain data directly collected from the HIUS as well as data derived from another source the Labour Force Survey (LFS). Demographic and employment data collected through the LFS were appended to the HIUS households. The LFS and HIUS data were collected from same households though not all households surveyed for the LFS were surveyed for the HIUS. For example, the total number of households surveyed for the HIUS 2003 is 23,113 while that for the LFS is 34,674. The data were collected through computer-assisted telephone surveys.

The data are available through the Tri-University Data Resources (TDR) website or through the new Nesstar web-site. The old website is no longer updated, though the full HIUS datasets of years 1997 to 2003 are available under the data group "Communications." Either a full set of observations can be downloaded or a sub-sample can be downloaded based on categories such as province, gender, etc. Furthermore, all variables (columns that match with questions in the questionnaire) or a subset of variables can be downloaded for each observation.

Estimation Results

We discuss the details of the estimations and results below. Table 1 in the Appendix A gives the descriptive statistics. The final number of observations from the pooled data of the years 1999, 2000, 2001, and 2003 that was used for the regressions is 48330 after about 1% of the observations had been discarded for reasons such as: one of the variables "value of orders" or "number of orders" had a non-zero value, but the other had a zero value; household had ordered a good in the "other – specify" category, so the category of good could not be ascertained; and both the variables "value of orders" and "number of orders" had non-zero value, but the household had no

[5] The data of 2002 were not used in the regressions due to format issues.
[6] Most of the content of this section is extracted from the HIUS 2003 User Guide published by Statistics Canada.

connection[7]. The mean value and number of orders is $ 266 and 2. Approximately 58% of the heads of households surveyed were in the 35 – 54 age group and 62% had completed high school. About 85% of the households surveyed had a male as the head of household. More than two-thirds of the households surveyed had annual income of at least C$ 44,000. About 76% of the households had only dial-up connection.

Step One Results

A variable was generated to indicate if a household had bought online or not. Using this as the dependent variable, a Probit regression and a Complementary Log-log (Cloglog) regressions were run. The Probit model pertains to the step one of the log value of orders model (continuous-type variable model) and the Cloglog regression pertains to the step one of the number of orders model (count-type variable model). It may be noted that the data are exactly the same for both the regressions since the generated variable "online buying or not" is the same either if generated from the value of orders data or from the number of orders data. Thus we run two different kinds of regressions – one Probit and the other Cloglog – on the same data for the probability of online buying. The former assumes that 0 arises when a normally distributed quantity linearly related to X, the vector of explanatory variables, is negative; whereas the complementary loglog regression assumes that the probability of 0 is Poisson with a log-linear link[8]. Table 2 in the Appendix B give the results[9]. We find that the results are qualitatively the same, though the point estimates are somewhat different. The results show that a household's probability of online buying is more when the head of household is in the youngest age group (< 35 years), is a male, his/her education is high, the household's income is high, and if it is connected through a high-speed connection. Both the regressions give the same qualitative results. We did not include product category variables in the first-step estimations since a household's first-ever online buying decision is probably independent of product nature. To such households, apprehension or ignorance about online buying may influence online buying decision.

Step Two Results

Households that cross the hurdle of the first-step face the step two decision of how much to buy or how frequently to buy online. These two different dependent variables are distributed as per a truncated normal distribution and a truncated Poisson distribution as already mentioned in the section on econometric models. Hence we run Truncated Regression and Truncated Poisson regression respectively with the above two dependent variables. Since the two-step model is advocated as superior to the one-step model it is sensible to compare the results of the two-step model with those of the one-step model. Hence here below we first discuss the results of the

[7] Maybe some households made online purchases from public computers connected to the Internet or from office computers, but there was no internet access at home.

[8] Katchova and Miranda (2004) do not appear to have estimated the complementary log-log model though they explain it. They report only their probit model estimates. Which of the two model assumptions for the probability of adopting is appropriate is not known, so we estimate both probit and complementary log-log models in this study.

[9] All the regressions for which results are presented were run without using weights. Statistics Canada cautions us to use weights, but we could not find software that run Tobit and Truncated Poisson regressions using sampling weight (which is the appropriate weight for survey data). However, the unweighted regression results and the weighted regression results of the Probit, Truncated Regression, Poisson, and Cloglog regressions (not presented here) were qualitatively the same; moreover, the unweighted results were qualitatively the same as the frequency weighted results of all the six regressions. We present the unweighted results so we can present the complete one-step and two-step estimation results. The very large number of observations may be helping us get robust results.

Truncated Regression in comparison with that of its corresponding (one-step) Tobit regression and then separately discuss the results of the Truncated Poisson regression in comparison with that of its corresponding (one-step) Poisson regression. We may here note that the step two regressions include product category variables, so we can get further insight about online buying decisions.

Truncated Regression vs. Tobit results

Table 3 in the Appendix C gives the results. The two models agree that: a household's online buying intensity (as measured by value of orders) is more when the education of its head of household is more, the household's income is more, and if its head is a male. But the two-step results suggest that the buying intensity is the highest if the head of household is in the 55 – 64 age group (highest for the youngest age group as per the one-step model), high if the household has high-speed connection (no difference between high-speed and dial-up as per the one-step model), and online buying is more for experience goods (more for search goods as per the one-step model).

The log-likelihood ratio (LR) test statistic is very high with associated probability nearly zero, so we have to reject the one-step Tobit model in preference to the two-step Truncated Regression. The step-one and step-two results together suggest that: the online buying probability and intensity (as measured by value of orders) are high if education and income are high, if head of household is a male, and for high-speed connection. While the probability of a household adopting online buying is more if its head is young, its intensity of buying is the highest if its head is in the 55 – 64 age group[10]. This is a notable result and provides evidence for the superiority of the two-step model over the one-step model that does not capture this result. The results also suggest that online buying is more for experience goods.

Truncated Poisson vs. Poisson regression results

Table 4 in the Appendix D gives the results. The two models agree that: a household's online buying intensity (as measured by number of orders) is more when the education of its head is more, income is more, if it has a high-speed connection, and if its head is a male. Both also agree that the ordering frequency is the highest for search goods. But the two models' results differ that: the ordering frequency is as high for the 55 – 64 age group as for the youngest age group as per the two-step model, whereas it is clearly the highest for the youngest age group as per the one-step model. In this case too, the LR statistic is very high with associated probability close to zero, so we must reject the one-step Poisson model in preference to the two-step Truncated Poisson model. The two-step results of the value of orders model and the number of orders model together suggest that value per order is the highest for experience goods (value is the highest, but frequency is the lowest for experience goods).

5. Conclusions

This study adopted a two-step estimation procedure to model probability of online buying in the first step and the intensity of online buying in the second step. The LR tests suggested rejecting the traditional one-step models of Tobit and Poisson regressions, used commonly with censored / truncated samples, in favor of Probit – Truncated Regression and Cloglog – Truncated Poisson regression two-step estimation procedures. Relaxing the restrictive assumption of the Tobit and Poisson models that the step one adoption decision process and the step two buying intensity

[10] It is possible that heads of households have a greater say in deciding whether to start buying online or not, but once started buying, factors such as family size, whether there are teen-agers in the family, etc. may influence buying intensity. This may explain why households in the 55 – 64 age group (among the online buying households) have the highest online buying intensity.

decision process are fundamentally one and the same leads to uncovering more insights in the two step procedures. Thus this study finds further empirical evidence in support of the two-step estimation procedure advocated by Katchova and Miranda (2004), so it may help increase the use of two-step procedure by various researchers.

The study also put the demographic variables, as predictors of online buying, in proper perspective. The use of a very large and high quality dataset from a reputed source and the use of superior estimation procedures are probably the reasons for getting unambiguous results with respect to the demographic variables. To sum up the results, the online buying probability is high when the head of household is young, a male, more educated, and when the household income is high and the household has a high-speed connection. The intensity of buying is the highest however for the 55 – 64 age group, though the other results for buying intensity are the same as those of buying probability.

The two-step results show, with respect to the product category variables, that value of orders is the highest for experience goods, but the ordering frequency is the lowest, so suggesting that dollar value per order is the highest for online purchases of experience goods. Regressions were also run with interaction variables. The results suggest that growth in buying of experience goods was the fastest over 1999-2003 in dollar terms; online buying of experience goods is low for lower age groups in dollar terms; online buying of experience goods is high for higher education and income groups in dollar terms; and online buying of credence goods is lower for male-headed households at least in ordering-frequency terms (not clear if it is so in dollar terms).

The study however has its own limitations. One such limitation is that its scope was not broadly aimed at understanding online buying behavior of households, only with respect to demographic variables and product category variables. It however showed, through its simple theoretical model, the complexity involved in comprehensively understanding online buying behavior. Another limitation is that the estimation procedure could be made further accurate through the use of sample weighting. It appears that various statistical software products in use are in continuing development, so do not have fully developed routines to handle not so common regressions like Truncated Poisson. The study had this limitation, but beyond its control.

Further research is clearly needed with respect to understanding consumer preferences and tastes. Further research could also attempt to study online buying with respect to product categories more thoroughly. A search good may be suitable to buy online (e.g. music) or not (e.g. shirts); an experience good may be suitable to buy online (e.g. hotel reservations) or not (e.g. food). Thus clearly there are further dimensions to product categorization, in the context of online buying, on top of the search-experience-credence theme. This offers scope for further research.

Appendix A

Table 1: Descriptive Statistics

Variable	Observations	Mean	Std. Dev.	Minimum	Maximum
Online-buying or Not[1]	48330	.3297124	.4701135	0	1
Value of Orders[2]	48330	266.2482	1264.695	0	100000
Value	48330	266.2482	1264.695	1.00e-09	100000
Log Value	48330	-12.05378	12.39027	-20.72327	11.51293
Number of Orders	48330	2.053217	7.539627	0	310
Age 1[3]	48330	.2092075	.4067471	0	1
Age 2	48330	.5814194	.4933314	0	1
Age 3	48330	.1372439	.3441083	0	1
Age 4	48330	.0721291	.2587042	0	1
Male	48330	.8534451	.3536653	0	1
Female	48330	.1465549	.3536653	0	1
Edu 1[4]	48330	.1196151	.3245143	0	1
Edu 2	48330	.6240223	.4843793	0	1
Edu 3	48330	.2563625	.4366288	0	1
Income 1[5]	48330	.1104283	.3134261	0	1
Income 2	48330	.2050279	.4037262	0	1
Income 3	48330	.3006414	.4585417	0	1
Income 4	48330	.3839023	.4863396	0	1
Dial-up	48330	.7594041	.4274498	0	1
High-speed	48330	.2405959	.4274498	0	1
Search	48330	.2717153	.4448485	0	1
Experience	48330	.1113387	.3145544	0	1
Credence	48330	.0111318	.1049196	0	1
Y99[6]	48330	.1872129	.3900864	0	1
Y00	48330	.2595282	.4383803	0	1
Y01	48330	.3146907	.4643974	0	1
Y03	48330	.2385682	.4262126	0	1

1 *Online-buying or Not* = 1 if buying online; 0 if not.
2 *Value of Orders* in C$; *Value* = *Value of Orders* + epsilon; and *Log Value* = ln(*Value*).
3 *Age 1*: < 35 yrs; *Age 2*: 35 – 54 yrs; *Age 3*: 55 – 64 yrs; and *Age 4*: 65 + yrs. Coded as indicator variables.
4 *Edu 1*: Less than high school; *Edu 2*: High school or some college; and *Edu 3*: University degree. Coded as indicator variables.
5 *Income 1*: = $ 24,000; *Income 2*: $ 24,001 - $ 43,999; *Income 3*: $ 44,000 - $ 69,999; and *Income 4*: $ 70,000 +. Coded as indicator variables.
6 *Y99*, *Y00*, *Y01*, and *Y03* are year controls for 1999, 2000, 2001, and 2003, respectively; year 2002 data were not used. Coded as indicator variables.

Appendix B

Table 2: Regression Results of the Step 1, of the Two-step Model, for Probability of Online Buying (Dependent Variable = *Online-buying or Not*)

Variable	Probit[1]	Complementary Log-Log
Constant	-1.297377[2]	-2.174025
	(.0324831) ***	(.0487486) ***
Age 2[3]	-.2210685	-.2885977
	(.015539) ***	(.0199163) ***
Age 3	-.2149073	-.2930193
	(.0211265) ***	(.0275509) ***
Age 4	-.3219333	-.4532212
	(.0273357) ***	(.0384751) ***
Female	-.0902337	-.1203073
	(.0185919) ***	(.0258778) ***
Edu 2	.2941633	.4394068
	(.0209421) ***	(.0319368) ***
Edu 3	.5829277	.8159836
	(.022934) ***	(.0336117) ***
Income 2	.158907	.2337998
	(.0243178) ***	(.0362458) ***
Income 3	.3082489	.441235
	(.0233564) ***	(.034448) ***
Income 4	.5216459	.7168312
	(.0235105) ***	(.0342773) ***
High-speed	.1597656	.204432
	(.0142505) ***	(.0181068) ***
Y00	.2568306	.3730214
	(.0191218) ***	(.0276399) ***
Y01	.389386	.5500099
	(.0184906) ***	(.026491) ***
Y03	.5752056	.7927361
	(.0195894) ***	(.0272666) ***
Number of observations	48330	48330
Zero outcomes		32395
Nonzero outcomes		15935
% correct predictions	68.67%	
Pseudo R^2[4]	0.0564	
Log likelihood	-28912.74	-28890.714
LR chi2[5]	3454.52	3498.57
Prob > chi2	0.0000	0.0000

1 Probit regression pertains to the Step 1 of the "Value of Orders" model that takes *Log Value* as the measure for online buying intensity; Complementary Log-Log regression pertains to the Step 1 of the "Number of Orders" model that takes *Number of Orders* as the measure.
2 Triple, double, and single asterisks denote significance levels of 1%, 5%, and 10%, respectively. Standard errors are in parentheses.
3 The omitted variables are *Age 1*, *Male*, *Edu 1*, *Income 1*, *Dial-up*, and *Y99*, respectively, for the sets of categorical variables pertaining to age, gender, education, income, type of internet connection, and year control; data for the year 2002 are not used.
4 "Pseudo R squared" is obtained by one minus the ratio of the full model's log-likelihood value to the constant-only model's log-likelihood value.
5 The chi-square test is to test the null hypothesis that all coefficients excluding constant are jointly zero.

Appendix C

Table 3: One-step vs. Two-step Regression Results for Intensity of Online Buying (Tobit[1] and Truncated Regressions)

Variable	Tobit (Dependent Variable = *Log Value*)	Truncated Regression (Dependent Variable = *Log Value*)
Constant	-35.61056[2]	3.750485
	(.3366134) ***	(.0739369) ***
Age 2[3]	-.6713483	.0010799
	(.1438768) ***	(.0267331)
Age 3	-.5477702	.0669322
	(.196737) ***	(.0368546) *
Age 4	-.6263615	-.0862313
	(.2606673) **	(.0518344) *
Female	-.2965967	-.1291203
	(.1771837) *	(.0343654) ***
Edu 2	1.08361	.109514
	(.2062854) ***	(.0433218) **
Edu 3	1.141495	.235387
	(.2232692) ***	(.0453899) ***
Income 2	.5535007	.0385261
	(.237221) **	(.0485999)
Income 3	1.212909	.1286898
	(.2266284) ***	(.0459857) ***
Income 4	1.340246	.3729667
	(.2278867) ***	(.0456895) ***
High-speed	.1589821	.0875719
	(.1312823)	(.0240608) ***
Search	33.29495	.6793054
	(.1435192) ***	(.0353225) ***
Experience	21.41839	1.223563
	(.1484937) ***	(.028504) ***
Credence	10.31426	.7458948
	(.388922) ***	(.0600501) ***
Y00	1.129927	.3658251
	(.1870256) ***	(.037162) ***
Y01	1.72608	.4229561
	(.1800832) ***	(.0355938) ***
Y03	2.150359	.7034865
	(.187973) ***	(.0364603) ***
Number of observations	48330	48330
Number of observations censored / truncated[4]	32395	32395
Number of observations uncensored / after truncation	15935	15935
Lower limit	-20.72326	-20.72326
Upper limit	+inf	+inf
Log likelihood	-63975.074	-27418.101
LR chi2[5]	62923.95	
Wald chi2		3077.93
Prob > chi2	0.0000	0.0000
Pseudo-R^2[6]	0.3297	
LR test for one-step vs. two-step model[7]	72767.2 (0.000)	

1 Tobit regression pertains to the one-step estimation of the "Value of Orders" model; Truncated regression pertains to the two-step estimation. The "Value of Orders" model takes *Log Value* as the measure for online buying intensity.
2 Triple, double, and single asterisks denote significance levels of 1%, 5%, and 10%, respectively. Standard errors are in parentheses.
3 The omitted variables are *Age 1*, *Male*, *Edu 1*, *Income 1*, *Dial-up*, and *Y99*, respectively, for the set of categorical variables pertaining to age, gender, education, income, type of internet connection, and year control; data for the year 2002 are not used.
4 Observations, for which the value of dependent variable <= the lower limit or >= the upper limit, are censored / truncated. These observations pertain to households that do not buy online.
5 The chi-square tests are to test the null hypothesis that all coefficients excluding constant are jointly zero.
6 "Pseudo R squared" is obtained by one minus the ratio of the full model's log-likelihood value to the constant-only model's log-likelihood value.
7 The likelihood-ratio test is given by LR = 2 *(ln L_{probit} + ln $L_{truncated\ regression}$ – ln L_{tobit}). The figure in the parentheses (in the table) is the associated chi-square probability. For the purpose of the LR test, the regressions of the two steps and that of the one-step model should have the same set of explanatory variables. The set of variables we used is that of the Probit / Complementary Log-log models.

Appendix D

Table 4: One-step vs. Two-step Regression Results for Intensity of Online Buying (Poisson[1] and Truncated Poisson Regressions)

Variable	Poisson Regression (Dependent Variable = *Number of Orders*)	Truncated Poisson Regression (Dependent Variable = *Number of Orders*)
Constant	-1.731053[2]	.3605691
	(.0223579) ***	(.0243626) ***
Age 2[3]	-.1065914	-.0740909
	(.0078457) ***	(.0079437) ***
Age 3	-.0366	-.0080205
	(.0106408) ***	(.010755)
Age 4	-.2081473	-.1834303
	(.0166248) ***	(.0169704) ***
Female	-.2726061	-.2721107
	(.0112717) ***	(.0115444) ***
Edu 2	.1347702	.0642044
	(.0136338) ***	(.0138648) ***
Edu 3	.1693265	.1018325
	(.0141985) ***	(.0143635) ***
Income 2	.1258082	.0902822
	(.0156889) ***	(.0160067) ***
Income 3	.1761632	.1083463
	(.0148393) ***	(.0151198) ***
Income 4	.2009643	.134637
	(.014707) ***	(.0149353) ***
High-speed	.1012941	.1030724
	(.0069246) ***	(.0069677) ***
Search	2.610118	.8503686
	(.0098115) ***	(.0107458) ***
Experience	1.068136	.6354824
	(.0066567) ***	(.0071811) ***
Credence	.7255317	.6606728
	(.013505) ***	(.0134614) ***
Y00	.415112	.3710654
	(.0124107) ***	(.012781) ***
Y01	.3854094	.3113564
	(.012082) ***	(.0124574) ***
Y03	.5264012	.4485744
	(.0121319) ***	(.012462) ***
Number of observations	48330	48330
Number of observations truncated[4]		32459
Number of observations after truncation		15935
Lower limit		0
Upper limit		+inf
Log likelihood	-110650.52	-89280.9719973
Model chi2[5]	192721.79	15194.83
Prob > chi2	0.0000	0.0000
Pseudo-R^2[6]	0.4655	0.0784
LR test for one-step vs. two-step model[7]	136229.63 (0.000)	

1 Poisson regression pertains to the one-step estimation of the "Number of Orders" model; Truncated Poisson regression pertains to the two-step estimation. The "Number of Orders" model takes *Number of Orders* as the measure for online buying intensity.
2 Triple, double, and single asterisks denote significance levels of 1%, 5%, and 10%, respectively. Standard errors are in parentheses.
3 The omitted variables are *Age 1*, *Male*, *Edu 1*, *Quart 1*, *Dial-up*, and *Y99*, respectively, for the set of categorical variables pertaining to age, gender, education, income, type of internet connection, and year control; data for the year 2002 are not used.
4 Observations, for which the value of dependent variable <= the lower limit or >= the upper limit, are truncated. These observations pertain to households that do not buy online.
5 The chi-square test is to test the null hypothesis that all coefficients excluding constant are jointly zero.
6 "Pseudo R squared" is obtained by one minus the ratio of the full model's log-likelihood value to the constant-only model's log-likelihood value.
7 The likelihood-ratio test is given by LR = 2(ln $L_{\text{complementary log-log}}$ + ln $L_{\text{truncated Poisson regression}}$ – ln L_{Poisson}). The figure in the parentheses (in the table) is the associated chi-square probability. For the purpose of the LR test, the regressions of the two steps and that of the one-step model should have the same set of explanatory variables. The set of variables we used is that of the Probit / Complementary Log-log models.

Appendix E
Categorization of Product Groups[11]

Search goods
Computer software
Computer hardware
Music (CDs, tapes, MP3)
Books, magazines, on-line newspapers
Videos, digital video disc (DVD)
Clothing, jewellery and accessories
Housewares (e.g. large appliances, furniture)
Consumer electronics (e.g. camera, computer, stereo, TV, VCR)
Automotive (cars, trucks, recreational vehicles or products)
Flowers – Gifts
Sports equipment
Toys and games
Real Estate
Other household related items
Banking

Experience goods
Other entertainment products (concert, theatre tickets)
Food, condiments, beverages
Travel arrangements (hotel reservations, travel tickets, rental car)
Crafts, hobbies, collectibles, antiques, art, garden, music instrument, pets
Other, Internet, renovations
Education
Internet on-line services
Antiques, collectibles and art

Credence goods
Health, beauty, medical, vitamins

[11] The product groups listed here are from those listed in the code books of the Household Internet Use Surveys of 1999 - 2003 (Public Use Micro-data Files) of Statistics Canada; categorization is ours as per Nelson (1970) and Darby & Karni (1973), based on dominant nature of goods.

References

Asch, D. 2001. “Competing in the New Economy.” *European Business Journal*; 2001; 13, 3; ABI / INFORM Global, pg. 119.

Bakos, Y. 1997. “Reducing Buyer Search Costs: Implications for Electronic Marketplaces.” *Management Science*, Vol. 43, No. 12 Frontier Research on Information Systems and Economics (Dec., 1997), pp. 1676 – 1692.

Bakos, Y. 2001. “The Emerging Landscape for Retail E-Commerce.” *The Journal of Economic Perspectives*, Vol. 15, No. 1 (Winter, 2001), pp. 69-80.

Bei, L., E.Y.I. Chen, and R. Widdows. 2004. “Consumers’ Online Information Search Behavior and the Phenomenon of Search vs. Experience Products.” *Journal of Family and Economic Issues*, 25(4): 449 – 467.

Bellman, S. Lohse, G. L., and Johnson, E. J. 1999. “Predictors of Online Buying Behavior.” *Communications of the ACM*, December 1999/Vol. 42, No. 12, p. 32

Borenstein S. and Saloner G. 2001. “Economics and E-commerce.” *The Journal of Economic Perspectives*, Vol. 15, No. 1 (Winter, 2001), pp. 3-12.

Darby, M.R. and Karni, E. 1973. “Free Competition and the Optimal Amount of Fraud.” *Journal of Law and Economics*, 16: 67-86.

Goolsbee, A. 2000. “In a world without borders: The impact of taxes on Internet Commerce.” *Quarterly Journal of Economics*, 115(2), 561-76.

Hans van der Heijden, Tibert Verhagen, and Marcel Creemers. 2001. “Predicting Online Purchase Behavior: Replications and Test of Competing Models.” Proceedings of the Hawai`i International Conference on System Sciences HICSS-34 January 3-6, 2001.

Hoffman, D. L. and Novak, T. P. 1999. “The Growing Digital Divide: Implications for an Open Research Agenda.” Retrieved August 13, 2006, from *http://199.239.233.76/downloadable_assets/digitaldivide_openresearch.pdf*

Katchova, A L. and Miranda, M. J. 2004. “Two-Step Econometric Estimation of Farm Characteristics Affecting Marketing Contract Decisions.” *American Journal of Agricultural Economics*. 86(1) (February 2004): 88 – 102.

Kline, L.R. 1998. “Evaluating the Potential of Interactive Media through a New Lens: Search versus Experience Goods.” *Journal of Business Research*, 41: 195-203.

Korgaonkar, P., R. Siverblatt, and T. Girard. 2006. “Online Retailing, Product Classifications, and Consumer Preferences.” *Internet Research*, 16(3): 267-288.

H. Li, C. Kuo, and M.G. Russel. 2003. "The Impact of Perceived Channel Utilities, Shopping Orientations, and Demographics on the Conumers' Online Buying Behavior" in *New Directions in Research on E-Commerce* by Charles Steinfield, Purdue University Press, ch4, p. 85.

Lohse, G. L., Bellman, S., and Johnson, E. J. 2000. "Consumer Buying Behavior on the Internet: Findings from Panel Data." *Journal of Interactive Marketing*; Winter 2000; 14, 1; ABI/INFORM Global, pg. 15.

Michalak, W. and Jones, K. 2003. "Canadian E-Commerce." *International Journal of Retail and Distribution Management*, Volume 31. Number 1. 2003. pp. 5-15.

Miyazaki, A. D. and Fernandez, A. 2001. "Consumer Perceptions of Privacy and Security Risks for Online Shopping." *Journal of Consumer Affairs*, SUMMER 2001 VOLUME 35, NUMBER 1 p27.

Nelson, P. 1970. "Information and Consumer Behavior." *Journal of Political Economy*, 78: 311-329.

Nelson P. 1974. "Advertising as Information." Journal of Political Economy. 83: 729 – 54.

Park, C. 2002. "A Model on the Online Buying Intention with Consumer Characteristics and Product Type." Proceedings of AusWeb 2002, The Eighth Australian World Wide Web Conference.

Tirole, J. 1989. "The Theory of Industrial Organization." MIT Press. 1988. Chapter 2.

Vulkan, N. 2003. "Automated E-Commerce" in *The Economics of E-Commerce: A Strategic Guide to Understanding and Designing the Online Marketplace.* Princeton University Press, Princeton, N J.

Wolfinbarger, M. and Gilly, M. 2001. "Shopping Online for Freedom, Control, and Fun." *California Management Review*. Berkeley: Winter 2001. Vol. 43, Iss. 2; p. 34 (22 pages)