DETERMINANTS OF ACCESS TO THE INTERNET IN HOUSEHOLDS – PROBIT MODEL ANALYSIS

HANNA DUDEK
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie

Summary

The purpose of this study was to analyse statistically factors that affect household access to Internet. The analysis was conducted using the Polish household budget survey 2003. There were a variety of factors influencing Internet usage. From the perspective of ecommerce, there were studied the linkages of economic, psychological and demographic variables to Internet use. There were taken into account many attributes of the household head and characteristics referring to whole household. The household access to the Internet was modelled using a probit estimation approach. In order to evaluate the validity of the estimation techniques, statistical tests were performed. Obtained results may help in creation of information society policy in Poland.

Keywords: households, Internet access, probit model

1. Introduction

The adoption of the Internet by Polish households in a little over a decade has been swift and impressive. Individuals have access to new technology is an essential part in transition towards the information society. The purpose of this study is to statistically analyse the factors that affect households’ access to internet. Households differ in size, age composition, educational level and other characteristics and, in general, we would expect households with different characteristics to having access to Internet at home. The results our research are based on probit regression analysis on Polish Household Budget Survey data which consists of 12732 records concerning of employees’ households in year 2003. In the probit analyses conducted for this study the research question is how region, household composition, age, gender, education, income and social position affect the probability of owning or having access to Internet at home. Binary choice models (probit or logit) allow estimation of, informally speaking, pure effects of household attributes, as the regression is run on all variables simultaneously.

2. Data

This paper was made possible by the availability of the Household Budget Survey (HBS) dataset collected by Polish Statistical Office in 2003. In the survey all the households are included, with the exception of those living in institutional households as well as households of members of the diplomatic corps of foreign countries. In the paper employees’ households were taken into account. In those households exclusive or main source of maintenance was income from hired work in either the public or private sector1. About 22 percent of households in that socio-economic

---

1 Household Budget Surveys in 2003
group had access to the Internet. As the aim of this research was statistical analysis of the determinants of households’ access to internet, many demographic, economic and sociologic variables were considered. In the literature a number of factors have repeatedly been identified and postulated as determinates of Internet use, most common postulated determinants are income, education, gender, ethnicity, and age\(^2\). They were taken into account in our research. The household access to the Internet was modelled using a probit estimation approach.

3. Probit models

Binary choice models have become very popular in applied researches. They allow to explore how each explanatory variable affects the probability of the event occurring. The event is captured by a dichotomous random variable \(y_i\) that takes two values, typically coded as 0 and 1. A general class of binary choice models has form:

\[
P_i = P(y_i = 1) = F(x_i\beta), \quad i = 1, 2, \ldots, n,
\]

where:
- \(P_i\) - a probability that \(y_i\) takes the value 1,
- \(\beta\) - a column vector of parameters \(\beta_1, \ldots, \beta_k\),
- \(x_i\) - a row vector representing the characteristics of individual \(i\),
- \(n\) - number of individuals,
- \(F\) is a CDF (Cumulative Distribution Function).

In applications usually following models are used:

- probit model with
  \[
  F(x, \beta) = \Phi(x\beta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt,
  \]
  (2)

- logit model with
  \[
  F(x, \beta) = \Lambda(x\beta) = \frac{1}{1 + \exp(-x\beta)}.
  \]
  (3)

Two models insure that the fitted values all lie inside the unit interval. From an empirical standpoint, it usually does not matter which model is applied. Logits and probits typically yield very similar results. This is because the cumulative distribution functions for the logit and probit are similar, differing slightly only in the tails of their respective distributions. In this paper probit model approach is implemented.

Since the probit model is nonlinear, the marginal effect in the outcome probability that is associated with a given change in the explanatory variable \(x_i\) depends on the levels of all independent variables:

\[
\frac{\partial P(y_i = 1|x)}{\partial x_i} = \beta_i \phi(x_i\beta),
\]

(4)

where \(\phi\) is the standard normal probability density function evaluated at \(x_i\beta\). Because a CDF is monotonically increasing in its argument, the second term in the chain rule derivative given above is always positive. As a result, the sign of the parameter \(\beta_i\) always equals the sign of the partial derivative of interest.

\(^2\) Genc I. H., Sahin H., Stone R. W. and Singh V.
The parameters of probit model we estimated by maximum likelihood method. We considered a following likelihood function:

\[ L = \prod_{i=1}^{n} P(y_i | x_i, \beta) = \prod_{i=1}^{n} \Phi(x_i \beta)^{y_i} (1 - \Phi(x_i \beta))^{1-y_i} \]  

(5)

Taking natural logarithms of \( L \) yields the log likelihood function for the probit model as:

\[ \ln L = \sum_{i=1}^{n} \left( y_i \ln \Phi(x_i \beta) + (1 - y_i) \ln(1 - \Phi(x_i \beta)) \right) \]  

(6)

Normal equations for such function are highly non-linear in their parameters and require solving by iterative methods (for example the Newton-Raphson method). Estimation of the probit model is straightforward using today’s powerful computers. Nowadays the procedures of maximum likelihood estimation of the probit model are implemented in many statistical software packages like Stata, PcGive, Statistica, SPSS and Gretl.

In order to evaluate the validity of the estimation techniques, statistical tests were performed. For testing a hypothesis about parameters, many procedures are available. The simplest method for a single restriction bases on t test. We can also test hypotheses that involve more that one restriction. For a test of the null hypothesis that all slope parameters are zero, the likelihood ratio test statistic is calculated as:

\[ LR = -2(\ln \hat{L}_{\text{Intercept}} - \ln \hat{L}_{\text{Full}}), \]  

(7)

where \( \ln \hat{L}_{\text{Intercept}} \) denotes the value of the restricted log-likelihood when all slope coefficients are zero and \( \ln \hat{L}_{\text{Full}} \) - the log-likelihood value for full model (without any restrictions imposed on parameters). Approximate critical values are obtained from the chi-square distribution with degrees of freedom equal to the number of slope parameters.

For binary choice models there are no firm ideas as to how to assess goodness-of-fit. The widely used McFadden’s proposal is to compare the model’s performance against a null model containing only an intercept:

\[ pseudo - R^2 = 1 - \frac{\ln \hat{L}_{\text{Full}}}{\ln \hat{L}_{\text{Intercept}}} \]  

(8)

4. Results

In our model dependent variable was defined as:

\[ y_i = \begin{cases} 
1, & \text{if } i\text{-th household had access to Internet at home} \\
0, & \text{otherwise} 
\end{cases} \]

\[ \hat{y}_i \]
There were taken into account many attributes of the households as explanatory variables. Table 1 presents the results of the multivariate probit analysis.

Table 1. Probit model estimation results

<table>
<thead>
<tr>
<th>Explanatory variable</th>
<th>Parameter estimate</th>
<th>t statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equivalent income using OECD scales</td>
<td>0.26</td>
<td>14.02</td>
</tr>
<tr>
<td>Edu_high (equals 1, if household’s head had higher or post-secondary education, 0 – otherwise)</td>
<td>1.18</td>
<td>12.92</td>
</tr>
<tr>
<td>Edu_sec (equals 1, if household’s head had only secondary education, 0 – otherwise)</td>
<td>0.74</td>
<td>8.88</td>
</tr>
<tr>
<td>Edu_voc (equals 1, if household’s head had only vocational education, 0 – otherwise)</td>
<td>0.40</td>
<td>4.69</td>
</tr>
<tr>
<td>Large_City (equals 1, if household lived in city with more than 500 thousand inhabitants, 0 – otherwise)</td>
<td>0.10</td>
<td>2.71</td>
</tr>
<tr>
<td>Small_City (equals 1, if household lived in city with 200-500 thousand inhabitants, 0 – otherwise)</td>
<td>0.11</td>
<td>2.61</td>
</tr>
<tr>
<td>Large_Town (equals 1, if household lived in town with 100-200 thousand inhabitants, 0 – otherwise)</td>
<td>0.13</td>
<td>2.77</td>
</tr>
<tr>
<td>Village (equals 1, if household lived in village, 0 – otherwise)</td>
<td>-0.24</td>
<td>-5.58</td>
</tr>
<tr>
<td>Age_Child13 (equals 1, if oldest child was above 13, 0 – otherwise)</td>
<td>0.67</td>
<td>20.91</td>
</tr>
<tr>
<td>Age_Child7 (equals 1, if oldest child was at least 7, 0 – otherwise)</td>
<td>0.35</td>
<td>8.88</td>
</tr>
<tr>
<td>Gender (equals 1, if household’s head was woman, 0 – otherwise)</td>
<td>-0.32</td>
<td>-10.56</td>
</tr>
<tr>
<td>Labour_Position (equals 1, if household’s head was employed in non-manual labour position, 0 – otherwise)</td>
<td>0.33</td>
<td>7.98</td>
</tr>
</tbody>
</table>

LR= 2262.16, LnL=-5613.98, Pseudo R²=0.17

Source: own calculations obtained by using Gretl Program

In order to evaluate the validity of the estimation techniques, statistical tests were performed. The likelihood ratio turns out to be 2262.16. With 12 degrees of freedom, the significance level associated with this statistic was below 0.05, concluding that a null hypothesis of insignificance of all slope coefficients was clearly rejected. From reported t-statistics we could state that all parameters were significant at 0.05. The value of maximum of logarithm of likelihood was -5613.98 and McFadden’s pseudo-R² - 0.17.

5. Conclusions

The empirical research presented in this paper tried to explain what drove the decision to have Internet at home. The carried out analysis under assumption ‘ceteris paribus’ has made possible to come to the following conclusions.

- Neither a age of household’s head nor a number of children was found to impact Internet access in households.
- The probability of having access to Internet at home was strongly affected by equivalent income. This result could encourage finding cost efficient ways to bring Internet. Cost of access to Internet and of equipment in relation to average earnings being in Poland one of the highest in Europe was the significant barrier.
The empirical results suggested that parenthood impacted on Internet use. Positive estimate of parameters occurs at variables referring to education. It denotes that a higher degree of educational attainment of household head was positive and significant predictor of Internet purchase. Clearly, human capital and family circumstances mattered a great deal to the determinants of net access at home.

The analysis showed that the bigger the city size, the higher the percentage of Internet access at home. Internet was more accessible in households living in the cities and large towns than households in small and medium-sized towns. The worse situation referred to rural households.

It was also found that in households with oldest child in school-age or older had were more likely to have net access than others. The Internet was widely used and had been annexed by the youngest generation of Poles as their means of communication. Children's and adolescent's access to IT could be expected to increase future employment and earning opportunities. This phenomenon can also be promise for the economic future: today’s young people who gain comfort and aptitude with new information technologies will be likely tomorrow’s skilled workers and innovators for our country.

Gender of household’s head seemed associated with a probability of choosing Internet access. Being female decreased the probability of having Internet at home. It could be explained by the fact that women were usually heads of household if they were alone – widowed or divorced. It denoted generally worse well-being of family.

Employment Status of Head of Household made differences in access of Internet. Households managed by person employed in non-manual labour position had generally more often Internet at home.

The results of the analyses done here provided interesting insights on the effects that characteristics of household had on the probability of having access to Internet. We realise that more research is needed on the potential consequences of differential access to information technology at home. Future study should also try to explain how policy could be better targeted to reduce the gap in the Internet access at home.

Bibliography

Hanna Dudek
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
Wydział Ekonomiczno-Rolniczy
Katedra Ekonometrii i Statystyki
Nowoursynowska 159
02-787 Warsaw, Poland
e-mail: hdudek@mors.sggw.waw.pl