Internet Usage Activity 3
More on Who is using the Internet?

Multinomial modeling, Chi-square

1 Introduction

At the end of 2001 the number of Internet users in the world was more than 500 million (up from 16 million in 1996). The Internet has quickly become part of our lives and numerous research efforts have been made in the past to try to understand who is using it and how it is being used. The activity presented here is concerned with those issues.

The data we use to that end comes from a survey conducted by the Graphics and Visualization Unit at Georgia Tech October 10 to November 16, 1997. The full details of this and other surveys conducted by that institution are available at http://www.gvu.gatech.edu/user_surveys/. Read through that page to get a feeling for what they found in previous surveys, other surveys and many other things.

The particular subset that we will analyze is the "general demographics" of internet users. We have recoded the data we downloaded from the survey web page as entirely numeric, with an index to codes described in http://kdd.ics.uci.edu/databases/internet_usage/changes.

The number of users participating in the survey is n= 10108.

2 Activities

Question 1:

The web page http://www.gvu.gatech.edu/user_surveys/ has a section called "Understanding how the Results are collected by reading" and under this section one called Survey Methodology. A statistician has to look at that page before looking at anything else. What kind of sampling, if any, was used to conduct the survey?Is it a probability sample? Explain. Can you generalize the results of this survey we are about to see to the overall population of WWW users? Why?

How could we improve Internet surveys if there is a need to improve them?

Question 2

You will need to answer this question in the lab opening the software package Stata. The commands given below are for Stata. IF you are rusty with Stata, you can visit `http://www.ats.ucla.edu/stat/stata`

For this question, you will need to read the data into Stata. Alternatively, you can read the text file that you can find in the same datasets directory by going to that directory with your browser.

Begin by loading the data into Stata:

· use `http://www.stat.ucla.edu/j̃sanchez/oid03/datasets/usage.dta`

Then we will label the values of income and code the missing values

· label define householdincome1 0 “not say” 1 “under $10” 2 “$10-19” 3 “$20-29” 4 “$30-39” 5 “$40-49” 6 “$50-74” 7 “$75-99” 8 “over $100”

· label value householdincome householdincome1

· mvdecode householdincome, mv(0)

After you have done this, tabulate the values of householdincome to obtain a table that will tell you the frequency and relative frequency of survey users in each income bracket.

What is this table telling you about the income level of survey participants? Are most survey participants in the upper income level, for example?

Question 3 Derive mathematically the maximum likelihood estimators for the parameters of the multinomial distribution for the number of people in each category of income (you do not need Stata for this, this is just a calculus problem).

Find numerically the value of the estimates of those parameters according to the survey data. You may or may not need Stata for this. You decide which is the case after you see the formulas you get.

Are the estimates you obtained good estimates of the true income distribution of all WWW users? Why?

Would the standard error of the estimates be meaningful?

Question 4:

According to the multinomial model that you just fitted to the data,what is the probability that there is an equal amount of users in each income category?

Question 5:

Find the income distribution of the United States in 1997, in a format identical to that used above for the distribution of income of the users. Compare the two distributions using a Chi-squared test. Is there statistical similarity between the two distributions? Does the usual interpretation of the Chi-square test follow here?

Question 6: The data you analyzed above is relatively old. We need to update the numbers. To do that, go to

`http://www.ccp.ucla.edu`

When there, check the “Findings from the First UCLA World Internet Report”, the findings about the “Use of the Internet by Latinos,” and the current release of the UCLA Internet report. Select from each of those information relevant to the one we analyzed here, i.e., income categories (internationally, for latinos and in general in the UCLA report). Compare that information to the one you obtained from the U. of Georgia survey. Are the summary statistics you found comparable to the ones of the Georgia survey?