

## Context

For a final xquery use case, I wanted a data situation where xquery would not normally be a tool of first choice. I wanted to use xquery where it was somewhat disadvantaged. So I chose survey research as the general area and the Current Population Survey as the specific instrument.

Normally, analysis of complex survey data would be done with a statistical tool like SAS or with a combination of a database (eg, relational), its query tool (eg, SQL), and a statistics package (eg, SAS or R). Based on previous xquery use cases I conducted, I felt reasonably confident that xquery would prove useful for survey research but that there might also be some rough edges. By comparing xquery with other tools I'd used previously, I figured I could experience xquery in a somewhat different way from the earlier use cases.

I also wanted to demonstrate an approach to survey analysis that I've found useful in the past. It is easy to generate reams of output when analyzing even moderately sized surveys. That's fine for a doctoral dissertation perhaps, but it's also the ultimate sleep enhancer for readers. Instead, I like to single out one area where survey respondents have made a decision. The CPS survey for October 2007 had a supplemental section devoted to school enrollment. So for demonstration purposes I'll look at the choices that first-year undergraduates made when deciding whether to attend public or private 4-year institutions.

In the United States, there is a very large gap in the cost of attendance between public and private 4-year colleges and universities. So the normal first assumption is that finances will be an important factor in these decisions. Let's find out (in a rough way). Things are not always what they seem. Hopefully I can provide some of the flavor of story-telling and analyst-as-detective while being true to the statistical minutia (it is important) but without hiding behind its facade.

## Summary

1. I continue to enjoy using xquery. Even under the mildly unfavorable conditions that I deliberately chose for this project, xquery proved useful for what Michael Driscoll likes to call data munging:

“[T]his refers to the painful process of cleaning, parsing, and proofing one's data before it's suitable for analysis. Real world data is messy. At best it's inconsistently delimited or packed into an unnecessarily complex XML schema. At worst, it's a series of scraped HTML pages or a thoroughly undocumented fixed-width format.”

Data munging is that and more. I like to call it living with the data until you know it. There are always idiosyncrasies and definitional nuances that can only be discovered by querying data, puzzling about the results, and repeating this process until you reach a high degree of comfort. This process often takes a long time unless you're lucky enough to have a mentor nearby who has already gone through the learning exercise with the data set.

I was quite happy with the ability of xquery to let me play with the CPS survey data and get to know it.

2. The results of the data analysis show a somewhat complex relationship between family income and student choice of a public or private college or university. To control the diversity of factors that affect this relationship, I looked only at students ages 15-19 who attended 4-year institutions full-time as first-year undergraduate students (ie, freshmen). Among this group, students whose families either reported the highest income levels, or who refused to report their income, were more likely to enroll at private institutions. But the relationship was modest and disappeared altogether after additional controls for employment status and race.

That may not seem like very exciting results, but getting to that point is instructive as to logic and method. So I urge you to read the section below on Analysis Results.

## Acknowledgements

Thanks to Jessica Davis and Hyon Shin of the Education and Social Stratification Branch of the U.S. Census Bureau for discussions about some of the definitions used in the October 2007 School Supplement to the Current Population Survey.

Thanks also to Daniela Florescu and Markos Zacharioudakis of zorba for discussions about xquery performance.

## Data Sources

The Current Population Survey (CPS) is a monthly survey of the civilian non-institutional population living in the United States. It is sponsored jointly by the U.S. Census Bureau and the U.S. Bureau of Labor Statistics and is the primary source for labor force statistics in the United States. The October survey includes the basic CPS and a set of supplemental questions on School Enrollment. The U.S.

Department of Education's National Center for Education Statistics jointly sponsors the supplemental school enrollment questions with the Census Bureau.

I used the CPS for October 2007 as the basis of this use case. Later data exists (October 2008) but I wanted a baseline prior to the time when the current recession started.

The links below provide more details on the CPS data used for this study.

October 2007 CPS data download  
GNU gzip, DOS/Windows, or Standard UNIX available  
151,370 record count  
each record contains 431 data variables in fixed format of total length 1056 characters  
[http://www.bls.census.gov/cps\\_ftp.html](http://www.bls.census.gov/cps_ftp.html)

Technical documentation (.pdf)  
October 2007: School Enrollment and Internet Use Supplement (10/07/2008)  
<http://www.census.gov/apsd/techdoc/cps/cpsoct07.pdf>

School Enrollment – Social and Economic Characteristics of Students, October 2007: Detailed Tables  
Table 5 is particularly useful for verification that the data is being read and weighted accurately.  
<http://www.census.gov/population/www/socdemo/school/cps2007.html>

Source and Accuracy Statement for the October 2007 CPS Microdata File on School Enrollment (.pdf)  
This is useful for understanding the weighting of sample results. It also has a good discussion on the sources of error possible in surveys and how these errors affect the accuracy of estimates.  
[http://www.census.gov/population/www/socdemo/school/2007school\\_S&A.pdf](http://www.census.gov/population/www/socdemo/school/2007school_S&A.pdf)

## General Procedures

1. Downloaded the October 2007 CPS data file (see link in section on Data Sources).
2. Used gawk to extract a subset of the variables of interest to me and to produce an initial flat xml file. The gawk program I wrote is available at: <http://garymlewis.com/instchg/public/xquery/cpsoct2007/cpsoct2007.awk>
3. Used xquery to create a data set to verify Table 5 in the detailed summary tables provided by the U.S. Bureau of Census (see link in section on Data Sources). This table refers to college students 15 years and older and shows distributions by age, sex, race, attendance status, control of school, and employment status. The xquery used to create the data is available at: [http://garymlewis.com/instchg/public/xquery/cpsoct2007/cpsoct2007\\_table5.xq](http://garymlewis.com/instchg/public/xquery/cpsoct2007/cpsoct2007_table5.xq)
4. Verified Table 5 to ensure that I was reading the data correctly with the gawk and xquery, and that I was applying the appropriate weights to individual records. One of several xquery programs I used is available at: [http://garymlewis.com/instchg/public/xquery/cpsoct2007/cpsoct2007\\_table5\\_1.xq](http://garymlewis.com/instchg/public/xquery/cpsoct2007/cpsoct2007_table5_1.xq)
5. Used xquery to explore various decisions that students make when deciding where to enroll. I chose to narrow the population of interest to enrolled students aged 15-19 in their first undergraduate year who attended 4-year institutions full time, and to examine various demographic and family characteristics related to the decision these students made about whether to attend public or private colleges and universities. An example of the xquery used at this stage is available at: [http://garymlewis.com/instchg/public/xquery/cpsoct2007/cpsoct2007\\_colluniv\\_explore\\_1.xq](http://garymlewis.com/instchg/public/xquery/cpsoct2007/cpsoct2007_colluniv_explore_1.xq)
6. Iterative exchanges between xquery and R until I got a data set useful for the analysis I wanted to conduct. The xquery used to define the final population for analysis and the variables to include is available at: [http://garymlewis.com/instchg/public/xquery/cpsoct2007/cpsoct2007\\_colluniv.xq](http://garymlewis.com/instchg/public/xquery/cpsoct2007/cpsoct2007_colluniv.xq)
7. Used R to do the final analysis. A history of the R is available at: [http://garymlewis.com/instchg/public/xquery/cpsoct2007/final\\_cpsoct2007\\_Rhistory.txt](http://garymlewis.com/instchg/public/xquery/cpsoct2007/final_cpsoct2007_Rhistory.txt)
8. The final R results (.pdf) are available at: [http://garymlewis.com/instchg/public/xquery/cpsoct2007/final\\_cpsoct2007\\_Rresults.pdf](http://garymlewis.com/instchg/public/xquery/cpsoct2007/final_cpsoct2007_Rresults.pdf)

## Analysis Results

Statistics can sometimes mislead. There's a now classic example with relevance to the analysis in the present project. Based on admissions to University of California Berkeley graduate schools in 1973, gender bias seemed evident. About 30% of women who applied to the six largest graduate programs at UC Berkeley were admitted, while the corresponding admit rate for men was nearly 45%. Flagrant if true. In fact, the numbers were incontrovertible. But further analysis showed the cause was not gender bias, but the simple fact that many more women than men applied to the graduate programs with the highest rejection rates. Once you examined admit rates by department, the gender differences disappeared.

Let's keep that in mind while using the October 2007 CPS survey to investigate whether family income affects students' choices about whether to enroll at a public or private college or university. In the United States there is a large difference in cost-of-attendance between public and private 4-year institutions. After considering financial aid awards and education tax benefits, the College Board estimates that the average net price for a full-time student attending a private 4-year not-for-profit institution was \$23,000 in 2007, compared to \$9,980 for a full-time student attending a public 4-year college or university. You'd expect that families with higher incomes would be better able to afford the higher net price at private institutions. For the College Board report *Trends in College Pricing 2007*, see: [http://www.trends-collegeboard.com/college\\_pricing/archive/CP\\_2007.pdf](http://www.trends-collegeboard.com/college_pricing/archive/CP_2007.pdf) and page 17 for the net price figures.

The October School Enrollment supplement to the CPS provides a nice opportunity to examine this assumption. There is a set of student choice questions and another set of demographic and family background antecedents that can help better understand enrollment choices. I included the following school choice issues: type of institution (4-year vs 2-year); attendance status (full-time or part-time); institutional control (public vs private); enrollment level (undergraduate vs graduate); and year of enrollment (first-year vs later). The individual and background factors I included were: sex; race (white non-Hispanic vs all others); Hispanic origin (no vs yes); age (15-19 vs 20+); employment status (employed vs not employed); and family income. The last variable deserves a bit more attention.

In the CPS, family income is recorded in 16 intervals. The lowest interval is "less than \$5,000" and the highest interval is "\$150,000 or more." Some respondents refuse to answer this question. Others don't know the answer. After considerable play with the family income variable, I grouped the highest income interval (\$150,000 or more) with those respondents who refused to answer the question. The analysis then compared this combined group with another group that contained all other income intervals. Whether this recode is defensible is open to debate, and I'll come back to the issue below.

All of these variables were recoded as binary (ie, 0 or 1) from variables in the CPS. If variables lend themselves to binary recodes, you can summarize relationships between variables in an easy-to-interpret fashion using correlation coefficients. These run from -1 through 0 to +1, with 0 meaning no correlation between two variables and both +1 and -1 meaning perfect correlation differing only in direction. Table 1 shows the correlations between the variables used in the analysis.

Here's a summary: Five of the correlations are significantly different from zero but even for these, the enduring emphasis is how modest these correlations are. Age does seem important. Older students are more likely to be employed while attending college or university; they are more apt to be minority (the word is used guardedly here to mean "other than white non-Hispanic"); and they are more likely to come from families with lower income levels. However the correlations are not strong. Our supposition in beginning this analysis was that family income would be associated with a greater probability of attending private institutions. That is apparently true based on results in Table 1, but the correlation is only 0.085. The fifth and final significant correlation in Table 5 is the one between employment status and income. Here, students who come from lower income families are more likely to be employed than those students from higher income families. There are probably no surprises in the direction of any of these 5 significant correlations.

**Table 1**

Correlations for Selected Variables in the October 2007 Current Population Survey:  
 Persons Enrolled Full-Time in their First Year at 4-Year Undergraduate Institutions

	Sex	Age	Empl	Race	Inc	Ctrl	Description	0	1
<b>Sex</b>	1.000	-0.017	-0.010	-0.014	-0.016	-0.010	Sex	Female	Male
<b>Age</b>		1.000	0.189	0.069	-0.080	-0.041	Age	15-19	>= 20
<b>Empl</b>			1.000	-0.026	-0.108	0.007	Employed?	No	Yes
<b>Race</b>				1.000	-0.060	-0.040	Race	White non-Hispanic	Other
<b>Inc</b>					1.000	0.085	Family Income	Other	>=150,000 +Refused
<b>Ctrl</b>						1.000	Control	Public	Private

Notes:

1. Gray cells indicate the correlation was significantly different from zero using a 95% confidence interval.

To better judge these correlations, it may help to look at them in a slightly different way. Table 2 shows the relationship between family income and public/private choice in a familiar tabular construction. It shows, for example, that the income distribution for students

attending private institutions is more skewed (38%) toward the higher income levels when compared to those attending public institutions (27%).

**Table 2**

Institution Control and Family Income in the October 2007 Current Population Survey:  
 Persons Enrolled Full-Time in their First Year at 4-Year Undergraduate Institutions

Family Income	Institution Control				Totals	
	Public		Private		Count	Percent
	Count	Percent	Count	Percent		
Other	1203	73%	243	62%	1446	71%
>= \$150,000 & Refused	434	27%	146	38%	580	29%
Totals	1637	100%	389	100%	2026	100%

Notes:

1. Counts are in thousands.
2. Gray cells highlight the apparent relationship between family income and enrollment at public or private institutions.
3. The correlation associated with this relationship is significant but a modest 0.085.

Let's pause for a moment and consider the correlations again. Even in this small group of variables, there appear to be interesting interactions between age, race, employment status, and income. It may be that our supposition about the relationship between family income and public/private institutional choice is similar to the UC/Berkeley example described earlier. Recall that the apparent relationship between graduate admissions rates and gender disappeared when a contravening variable (academic department) was introduced.

At this point, an analyst would probably begin to play with models to better understand what's happening in the data. A simple model that could explain the correlations in Table 1 goes like this: family income does seem to affect public/private decisions but there are antecedents in the form of intermingled age, race, and employment status that affect income. Controlling for these intermingled variables may be important for understanding what the data is actually saying.

I won't take you through any other steps in the analysis. But here's a summary. Age seems central, so I first controlled for age by restricting the students to only those in the 15-19 year category. The strength of the relationship between family income and public/private choice actually increased a bit to 0.093. But so too does the strength of the relationship between family income and one of the other two important variables (employment status). So we still likely have some confounding occurring, and controlling for another variable seems warranted. And so the story goes. The end point is that if you control for age, race, and employment status, the relationship between family income and public/private choice nearly disappears.

However, the danger of applying deeper levels of control on the data is that the numbers of people included become smaller and smaller. In fact, the Census Bureau warns that for the October 2007 CPS "summary measures (such as medians and percentage distributions) probably do not reveal useful information when computed on a subpopulation smaller than 75,000." Controlling the relationship between family income and public/private choice by age, race, and employment status approaches this 75,000 level in some of the table cells.

What to make of all this? Well, it nicely demonstrates yet again that correlations can be devilishly tricky to interpret and require great care in analysis. Second, it also raises the possibility that my choice for recoding family income is invalid. I would not care to defend my choice, but this is a demonstration project and I took some latitude that I likely would not have otherwise. Third, the population was constrained in a peculiar way right from the start. I was examining the relationship between family income and public/private choice among students who had all decided to matriculate. We might see a stronger influence exerted by family income if we considered the actual enrollment decision (ie, who attends higher education vs who does not?). And, fourth, the family financial data collected in the CPS is not as strong one might hope. For example, there is no asset question comparable to the one on income. The strength of family finances is really a combination of income and assets.