THIRD EDITION

PUBLIC FINANCE AND PUBLIC POLICY

JONATHAN GRUBER
Massachusetts Institute of Technology

Worth Publishers

Appendix to Chapter 3

Cross-Sectional Regression Analysis

In the text, we presented a cursory discussion of cross-sectional regression analysis, and the role of control variables. In this appendix, we provide a more detailed presentation of this approach within our TANF example.

Data For this analysis, we use data from the March 2002 Current Population Survey (CPS). From that survey, we selected all women who reported that they were unmarried and had a child younger than age 19. The total sample is 8,024 single mothers.

For this sample, we have gathered data on the following variables for each woman:

- *TANF:* Total cash TANF benefits in the previous year (in thousands of dollars).
- *Hours:* Total hours of work in the previous year, computed as reported weeks of work times usual hours per week.
- *Race:* We divide reported race into white, black, and other.
- *Age:* Age in years.
- *Education:* We use reported education to divide individuals into four groups: high school dropouts; high school graduates with no college; those with some college; and college graduates.
- *Urbanicity:* We use information on residential location to divide individuals into four groups: central city; other urban; rural; and unclear (the CPS doesn't identify location for some mothers for survey confidentiality reasons).

Regression Using these data, we can estimate a regression of the impact of welfare on hours of work of the form:

$$(1)\ \text{HOURS}_i = \alpha + \beta \text{TANF}_i + \epsilon_i$$

where there is one observation for each mother i. This is the counterpart of the regression analysis shown in Figure 3-4, but now we are using each individual data point, rather than grouping the data into categories for convenience.

In this regression, α, the constant term, represents the estimated number of hours worked if welfare benefits are zero. β is the slope coefficient, which

represents the change in hours worked per dollar of welfare benefits. ϵ is the error term, which represents the difference for each observation between its actual value and its predicted value based on the model.

The results of estimating this regression model are presented in the first column of the appendix table. The first row shows the constant term α, which is 1,537: this measures the predicted hours of labor supply delivered at zero welfare benefits. The second row shows the coefficient β, which is -107: each \$1,000 of welfare benefits lowers hours worked by 107. This is very close to the estimate from the grouped data of -110 discussed in the text. Thus, for a mother with no welfare benefits, predicted hours of work are 1,537; for a mother with \$5,000 in welfare benefits, predicted hours of work are $1{,}537 - 5 \times 107 = 1{,}002$.

Underneath this estimate in parentheses is the estimate's *standard error*. This figure captures the precision with which these coefficients are estimated and reminds us that we have here only a statistical representation of the relationship between welfare benefits and hours worked. Roughly speaking, we cannot statistically distinguish values of β that are two standard errors below or above the estimated coefficient. In our context, with a standard error of 3.7 hours, the results show that our best estimate is that each thousand dollars of welfare lowers hours worked by 107, but we can't rule out that the effect is only 99.6 ($107 - 2 \times 3.7$) or that it is 114.4 ($107 + 2 \times 3.7$).

In the context of empirical economics, this is a *very* precise estimate. Typically, as long as the estimate is more than twice the size of its standard error, we say that it is *statistically significant*.

The final row of the table shows the R^2 of the regression. This is a measure of how well the statistical regression model is fitting the underlying data. An R^2 of 1 would mean that the data are perfectly explained by the model so that all data points lie directly on the regression line; an R^2 of 0 means that the data are not at all explained. The value of 0.095 here says that less than 10% of the variation in the data is explained by this regression model.

As discussed in the text, however, this regression model suffers from serious bias problems, since those mothers who have a high taste for leisure will have both low hours of work and high welfare payments. One approach to addressing this problem suggested in the text was to include control variables. We don't have the ideal control variable, which is taste for leisure. We do, however, have other variables that might be correlated with tastes for leisure or other

■ APPENDIX 3 TABLE

Cross-Sectional Regression Analysis

	Equation (1)	Equation (2)
Constant	1,537	2,062
	(10)	(61)
TANF benefits	−107	−93
	(3.7)	(3.6)
White		181
		(44)
Black		61
		(47)
High school dropout		−756
		(30)
High school graduate		−347
		(25)
Some college		−232
		(28)
Age		−9.3
		(0.8)
Central city		−12
		(30)
Other urban		34
		(29)
Rural		−43
		(31)
R^2	0.095	0.183

factors that determine labor supply: race, education, age, and urbanicity. So we can estimate regression models of the form:

$$(2)\ \text{HOURS}_i = \alpha + \beta\text{TANF}_i + \delta\text{CONTROL}_i + \epsilon_i$$

where CONTROL is the set of control variables for individual *i*.

In the second column of the appendix table, we show the impact of including these other variables. When we have a categorical variable such as race (categorized into white, black, and other), we include *indicator variables* that take on a value of 1 if the individual is of that race, and 0 otherwise. Note that when we have N categories for any variable (e.g., 3 categories for race), we only include $N - 1$ indicator variables, so that all estimates are relative to the excluded category (e.g., the coefficient on the indicator for "black" shows the impact of being black on welfare income, relative to the omitted group of Hispanics).

Adding these control variables does indeed lower the estimated impact of welfare benefits on labor supply. The coefficient falls to –93, but remains highly significant. The R^2 doubles but still indicates that we are explaining less than 20% of the variation in the data.

The control variables are themselves also of interest:

- *Race:* Whites are estimated to work 181 hours per year more than Hispanics (the omitted group); blacks are estimated to work 61 hours per year more than Hispanics, but this estimate is only about 1.3 times as large as its standard error, so we do not call this a statistically significant difference.
- *Education:* Hours of work clearly rise with education. High school dropouts work 756 fewer hours per year than do college graduates (the omitted group); high school graduates work 347 fewer hours per year; and those with some college work 232 fewer hours per year than those who graduate from college. All of these estimates are very precise (the coefficients are very large relative to the standard errors beneath them in parentheses).
- *Age:* Hours worked decline with age, with each year of age leading to 9 fewer hours of work; this is a very precise estimate as well.
- *Location:* Relative to those with unidentified urbanicity, people in cities and rural areas work less and those in the suburbs work more, but none of these estimates is statistically precise.

Do these control variables eliminate bias in the estimated relationship between TANF benefits and labor supply? There is no way to know for sure, but it seems unlikely. The fact that this large set of controls explains only 9% more of the variation in labor supply across individuals suggests that it is unlikely to capture all of the factors correlated with both labor supply and TANF benefits.