

Very Short Paper I: Data Collection
due: Wednesday Oct 11

Go find some interesting data.

Your data should include at least 50 observations of values for at least two variables in a cross-section or a panel ($N \geq 50$, $T > 0$).

Report a table of means, which includes for each variable to be used in the analysis, the mean, standard deviation, number of observations, minimum and maximum value (i.e., `summi` in Stata^{*}). If weights are to be used in analysis, report the weighted means and standard deviations as well. At the bottom of the table you should document the source, including enough information to make replication possible.

The very short paper will be marked on the econometric method alone, with no marks deducted for even the most ludicrous economic analysis, so feel free to have fun. On the other hand, you will spend many intimate hours with this project, so you may as well construct it in a way that will make it interesting for yourself.

Feel free to contact any of us (by Oct 2) if you have questions about data choices.

Note: This example includes the number of obs in the note. The min and max are often excluded in the published version of a paper due to space constraints. A good first draft should have them.

TABLE 3.—DESCRIPTIVE STATISTICS FOR OIL REFINERIES PACE, LRD, AND REGULATORY DATA

	Mean	Standard Deviation
Value of shipments*	1,707,848	2,890,197
Value added	118,772	231,349
Employment	372	500
Air pollution abatement investment	2,096	7,618
Net abatement investment	1,495	7,475
Depreciation of abatement capital	601	1,796
Abatement operating costs	6,586	16,607
Change in abatement operating costs	141	6,951
New regulation adoption dates	0.053	0.369
New regulation compliance dates	0.041	0.267
New increased stringency dates	0.012	0.136
South coast indicator	0.055	0.228
California indicator	0.129	0.335
Texas indicator	0.208	0.406
Louisiana indicator	0.094	0.292

^{*} Thousands of 1991 dollars deflated by the Producer Price Index.

Source: Pollution Abatement Costs and Expenditure microdata.

The sample contains 1,914 observations weighted by PACE sampling weights to represent 2,425 plant-years in the population. Sampled from 1979–1991, excluding 1983 and 1987. Data from 1992 and 1993 were excluded due to errors. Change in operating costs is from year to year and is defined only for plants observed for two consecutive sampled years. Employment is measured in persons.

^{*} If you are thinking of using some software other than Stata please contact one of us so we can arrange for an MRI.