# DATA SET HANDBOOK

## Introductory Econometrics: A Modern Approach, 2e
## Jeffrey M. Wooldridge

This document contains a listing of all data sets that are provided with the second edition of *Introductory Econometrics: A Modern Approach*. For each data set, I list its source (wherever possible), where it is used or mentioned in the text (if it is), and, in some cases, notes on how an instructor might use the data set to generate new homework exercises, exam problems, or term projects; in some cases, I suggest ways to improve the data sets. Occasionally, I will update the document to provide new ideas for how to use the data sets.

**401K.RAW**

**Source:** L.E. Papke (1995), "Participation in and Contributions to 401(k) Pension Plans: Evidence from Plan Data," *Journal of Human Resources* 30, 311-325.

Professor Papke kindly provided these data. She gathered them from the Internal Revenue Service's Form 5500 tapes.

**Used in Text:** pages 64-65, 80, 134-135, 171-172, 213, 663-665

**Notes:** This data set is used in a variety of ways in the text. One additional possibility is to investigate whether the regression functions of *prate* on *mrate*, and the firm size variables differ by whether the plan is a sole plan. The Chow test, and the variant that allows different intercepts, can be used.

**401KSUBS.RAW**

**Source:** A. Abadie (2000), "Semiparametric Estimation of Instrumental Variable Models for Causal Effects," NBER Technical Working Paper No. 260.

Professor Abadie kindly provided these data. He obtained them from the 1991 Survey of Income and Program Participation (SIPP).

**Used in Text:** pages 165, 255, 256, 287-288, 321, 521

**Notes:** This data set can also be used to illustrate the nonlinear binary response models in Chapter 17, where, say, *pira* is the dependent variable, and *e401k* is the key independent variable, in a probit or logit model.

**ADMNREV.RAW**

**Source:** Data from the National Highway Traffic Safety Administration: "A Digest of State Alcohol-Highway Safety Related Legislation," U.S. Department of Transportation, NHTSA. The third (1985), eighth (1990), and 13th (1995) editions were used.

**Used in Text:** not used

**Notes:** This is not so much a data set as a summary of so-called "administrative per se" laws at the state level, for three different years. It could be supplemented with drunk driving fatalities for a nice econometric analysis. In addition, the data for 2000 can be added. It could form the basis for a term project. Many other explanatory variables could be included. Unemployment rates, State-level tax rates on alcohol, and membership in MADD, are just a few possibilities.

**AFFAIRS.RAW**

**Source:** R.C. Fair (1978), "A Theory of Extramarital Affairs," *Journal of Political Economy* 86, 45-61, 1978.

I collected the data from Professor Fair's web cite at the economics department at Yale University. He originally obtained the data from a survey by *Psychology Today*.

**Used in Text:** not used

**Notes:** This would make an interesting data set for problem sets, starting from Chapter 7. Even though *naffairs* is a count variable, a linear model can be used. Or, you could ask the students to estimate a linear probability model for *affair*. One possibility is to test whether putting the marriage rating variable, *ratemarr*, is enough, against the alternative that a full set of dummy variables is needed; see page 229 for a similar example. This is also a good data set to illustrate Poisson regression, or probit and logit, in Chapter 17.

**AIRFARE.RAW**

**Source:** Jiyoung Kwon, a doctoral candidate in economics at MSU, kindly provided these data, which she obtained from the Domestic Airline Fares Consumer Report by the U.S. Department of Transportation. The web site is http://ostpxweb.ost.dot.gov/aviation/.

**Used in Text:** not used

**Notes:** The report cited above provided information about average prices being paid by consumers in the top 1000 largest domestic city-pair markets within the 48 contiguous states. These markets account for about 75 percent of all 48-state passengers and 70 percent of total domestic passengers. The data in this paper include the top 1000 city-pair markets for each fourth quarter of 1997 to 2000. This is a large panel data set that can nicely illustrate the different results that can be obtained from pooled OLS, random effects, and fixed effects. The dependent variable can be *fare* or, even better, its natural log. The key explanatory variable is the market share of the largest carrier. The route distance should be included as well.

An interesting possibility is to estimate a demand function, where log(*passen*) is the dependent variable, log(*fare*) is the potentially endogenous explanatory variable, and log(*dist*) and its square are other factors affecting demand. If you estimate this equation by OLS using, say, the latest year (2000), you get a negative fare elasticity. If you instead use *concen* as an IV for log(*fare*) – so the assumption is that concentration affects the fare but not the demand on the route – then the elasticity is much larger.

**APPLE.RAW**

**Source:** These data were used in the doctoral dissertation of Jeffrey Blend, Department of Agricultural Economics, Michigan State University, 1998. The thesis was supervised by Professor Eileen van Ravensway. Drs. Blend and van Ravensway kindly provided the data. The data come from a telephone survey conducted by the Institute for Public Policy and Social Research at MSU.

**Used in Text:** pages 597-598

**Notes:** While these data are not used until a problem in Chapter 17, they can be used much earlier in a linear regression model to illustrate estimation of an economic model with truly exogenous variables – the price variables, in this case. This is the closest thing to experimental data that I have. The own price effect is strongly negative, the cross price effect is strongly positive. Interestingly, because the survey design induces a strong positive correlation between the prices of eco-labeled and ordinary apples, there is an omitted variable problem if either is dropped from the demand equation. A good exam question is to show a simple regression of *ecolbs* on *ecolbs* and then a multiple regression on both prices, and ask students to decide whether the price variables are positively or negatively correlated.

**ATHLET1.RAW**

**Sources:** *Peterson's Guide to Four Year Colleges*, 1994 and 1995 (24th and 25th editions). Princeton University Press. Princeton, NJ.

*The Official 1995 College Basketball Records Book*, 1994, NCAA.

*1995 Information Please Sports Almanac* (6th edition). Houghton Mifflin. New York, NY.

**Used in Text:** page 669

**Notes:** These data were collected by Patrick Tulloch, a former MSU undergraduate, for a term project. The "athletic success" variables are for the year prior to the enrollment and academic data. Updating these data to get a longer stretch of years, and including appearances in the "Sweet 16" NCAA basketball tournaments, would make for a more convincing analysis. With the growing popularity of women's sports, especially basketball, an analysis that includes success in Women's athletics would be interesting.

**ATHLET2.RAW**

**Sources:** *Peterson's Guide to Four Year Colleges*, 1995 (25th edition). Princeton University Press.

*1995 Information Please Sports Almanac* (6th edition). Houghton Mifflin. New York, NY

**Used in Text:** page 669

**Notes:** These data were collected by Paul Anderson, a former MSU undergraduate, for a term project. The score from football outcomes for natural rivals (Michigan-Michigan State, California-Stanford, Florida-Florida State, to name a few) is matched with application and academic data. The application and tuition data are for Fall 1994. Football records and scores are from 1993 football season.

**ATTEND.RAW**

**Source:** These data were collected by Professors Ronald Fisher and Carl Liedholm during a term in which they both taught principles of microeconomics at Michigan State University. Professors Fisher and Liedholm kindly gave me permission to use a random subset of their data, and their research assistant at the time, Jeffrey Guilfoyle, provided helpful hints.

**Used in Text:** pages 112, 151, 195-196, 213, 215-216

**Notes:** The attendance figures were obtained by requiring students to slide their ID cards through a magnetic card reader, under the supervision of a teaching assistant. You might have the students use *final*, rather than the standardized variable, so that they can see the statistical significance of each variable remains exactly the same. The standardized variable is used only so that the coefficients measure effects in terms of standard deviations from the average score.

**AUDIT.RAW**

**Source:** These data come from a 1988 Urban Institute audit study in the Washington, D.C. area. I obtained them from the article "The Urban Institute Audit Studies: Their Methods and Findings," by James J. Heckman and Peter Siegelman. In Fix, M. and Struyk, R., eds., *Clear and Convincing Evidence: Measurement of Discrimination in America.* Washington, D.C.: Urban Institute Press, 1993, 187-258.

**Used in Text:** pages 755-756, 762, 766

**BARIUM.RAW**

**Source:** C.M. Krupp and P.S. Pollard (1999), "Market Responses to Antidumpting Laws: Some Evidence from the U.S. Chemical Industry," *Canadian Journal of Economics* 29, 199-227.

Professor Krupp kindly provided the data. They are monthly data covering February 1978 through December 1988.

**Used in Text:** pages 342-343, 354, 356, 357, 401, 405-406, 422, 633, 635, 643

**Notes:** Rather than just having intercept shifts for the different regimes, one could conduct a full Chow test across the different regimes.

**BWGHT.RAW**

**Source:** J. Mullahy (1997), "Instrumental-Variable Estimation of Count Data Models: Applications to Models of Cigarette Smoking Behavior," *Review of Economics and Statistics* 79, 596-593.

Professor Mullahy kindly provided the data. He obtained them from the 1988 National Health Interview Survey.

**Used in Text:** pages 63, 110, 150-151, 164, 174, 180-181, 182-184, 249-250, 494

**BWGHT2.RAW**

**Source:** Zhehui Luo, a doctoral candidate in economics at MSU, kindly provided these data, which she obtained from state files linking birth and infant death certificates, and from the National Center for Health Statistics natality and mortality data.

**Used in Text:** pages 216-217

**Notes:** Much can be done with this data set. In addition to number of prenatal visits, smoking and alcohol consumption (during pregnancy) are included as explanatory variables. These can be added to equations of the kind found in Computer Exercise 6.17. In addition, the one- and five-minute APGAR scores are included. These are measures of the well-being of infants just after birth. An interesting feature of the score is that it is bounded between zero and 10, making a linear model less than ideal. Still, a linear model would be informative, and you can ask students about predicted values less than zero or greater than 10.

**CAMPUS.RAW**

**Source:** These data were collected by Daniel Martin, a former MSU undergraduate, for a final project. They come from the FBI Uniform Crime Reports and are for the year 1992.

**Used in Text:** pages 129-130

**Notes:** Colleges and Universities are now required to provide much better, more detailed crime data. A very rich data set can now be obtained, even a panel data set for colleges across different years. Statistics on male/female ratios, fraction of men/women in fraternities or sororities, particular policy variables – such as a "safe house" for women on campus – could be added as explanatory variables. The crime rate in the host town would also be a good control.

**CARD.RAW**

**Source:** D. Card (1995), "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp. Ed. L.N. Christophides, E.K. Grant, and R. Swidinsky, 201-222. Toronto: University of Toronto Press.

Professor Card kindly provided these data.

**Used in Text:** pages 497-499, 519, 520

**Notes:** Computer Exercise 15.14 is important for analyzing these data. There, it is shown that the instrumental variable, *nearc4*, is actually correlated with *IQ*, at least for the subset of men for which an IQ score is reported. However, the partial correlation between *nearc4* and *IQ* is arguably zero, or at least not statistically different from zero.

For a more advanced course, a nice extension of Card's analysis is to allow the return to education to differ by race. A simple analysis includes *black·educ* as an additional explanatory variable, and uses as its instrument *black·nearc4*.

**CEMENT.RAW:**

**Source:** J. Shea 1993), "The Input-Output Approach to Instrument Selection," *Journal of Business and Economic Statistics* 11, 145-156.

Professor Shea kindly provided these data.

**Used in Text:** page 551

**Notes:** Compared with Shea's analysis, the producer price index (PPI) for fuels and power has been replaced with the PPI for petroleum. The data are monthly and have note been seasonally adjusted.

**CEOSAL1.RAW**

**Source:** I took a random sample of data reported in the May 6, 1991 issue of *Businessweek*.

**Used in Text:** pages 33, 41, 158-159, 200, 211, 251, 254, 320, 670

**Notes:** This kind of data collection is very easy for students just learning data analysis, and the findings are sometimes interesting. A good term project is to have students collect a similar data set using a more recent issue of *Businessweek*.

**CEOSAL2.RAW**

**Source:** See CEOSAL1.RAW

**Used in Text:** pages 65, 110-111, 162-163, 209-210, 318, 670

**Notes:** In this version of the data set, more information about the CEO, rather than about the company, is included.

**CONSUMP.RAW**

**Source:** I collected these data from the 1997 *Economic Report of the President*. Specifically, the data come from Tables B-71, B-15, B-29, and B-32.

**Used in Text:** pages 358, 389, 422, 542-543, 551, 644

**Notes:** For a student interested in time series methods, updating this data set and using it in a manner similar to that in the text could be acceptable as a final project.

**CORN.RAW**

**Source:** G.E. Battese, R.M. Harter, and W.A. Fuller (1988), "An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data," *Journal of the American Statistical Association* 83, 28-36.

This small data set is reported in the article.

**Used in Text:** pages 771-772

**Notes:** You could use these data to illustrate simple regression, where the intercept should be zero: no corn pixels should predict no corn planted.

**CPS78_85.RAW**

**Source:** Professor Henry Farber, now at Princeton University, compiled these data from the 1978 and 1985 Current Population Surveys. Professor Farber kindly provided these data when we were colleagues at MIT.

**Used in Text:** pages 430-431, 455

**Notes:** Obtaining more recent data from the CPS allows one to track, over a long period of time, the changes in the return to education, the gender gap, black-white wage differentials, and the union wage premium.

**CPS91.RAW**

**Source:** Professor Daniel Hamermesh, now at the University of Texas, compiled these data from the May 1991 Current Population Surveys. Professor Hamermesh kindly provided these data.

**Used in Text:** not used

**Notes:** This is much bigger than the other CPS data sets, and it is much bigger than MROZ.RAW, too. So, this data set can be used in a standard wage equation for married women. We also have information on the husband, and so a labor supply function can be estimated as in Chapter 16, although the validity of potential experience as an IV for log(*wage*) could be questioned. (MROZ.RAW contains an actual experience variable.) Perhaps more convincing is to add hours to the wage offer equation, and instrument hours with number of young children and number of old children. This data set also contains a union indicator.

**CRIME1.RAW**

**Source:** J. Grogger (1991), "Certainty vs. Severity of Punishment," *Economic Inquiry* 29, 297-309.

Professor Grogger kindly provided a subset of the data he used in his article.

**Used in Text:** pages 82-83, 171, 177, 244-245, 264, 286, 290-292, 576-578, 596

**CRIME2.RAW**

**Source:** These data were collected by David Dicicco, a former MSU undergraduate, for a final project. They came from various issues of the *County and City Data Book*, and are for the years 1982 and 1987. Unfortunately, I do not have the list of cities.

**Used in Text:** pages 300-301, 438-441

**Notes:** Very rich crime data sets, at the county or even the city level, can be collected using the FBI's *Uniform Crime Reports*. These can be matched up with demographic and economic data, at least for census years. The *County and City Data Book* usually contains a variety of statistics, but the years do not always match up.

**CRIME3.RAW:**

**Source:** E. Eide (1994), *Economics of Crime: Deterrence of the Rational Offender*. Amsterdam: North Holland. The data come from Tables A3 and A6.

**Used in Text:** pages 443-444, 456

**Notes:** These data are for the years 1972 and 1978 for 53 police districts in Norway. Much larger data for more years can be obtained for the United States, although a measure of the "clear-up" rate is needed.

**CRIME4.RAW**

**Source:** From C. Cornwell and W. Trumball (1994), "Estimating the Economic Model of Crime with Panel Data," *Review of Economics and Statistics* 76, 360-366.

Professor Cornwell kindly provided the data.

**Used in Text:** pages 451-452, 457, 478, 451-452, 457, 478, 551-552

**Notes:** Computer Exercise 16.15 shows that variables that might seem to be good instrumental variable candidates are not always so good, especially after a transformation such as differencing across time. You could have the students do an IV analysis for just, say, 1987.

**DISCRIM.RAW**

**Source:** K. Graddy (1997), "Do Fast-Food Chains Price Discriminate on the Race and Income Characteristics of an Area?" *Journal of Business and Economic Statistics* 15, 391-401.

Professor Graddy kindly provided the data.

**Used in Text:** page 671

**Notes:** This data set is actually only mentioned in the text; it is not used in examples or exercises. If you want to assign a common final project, this would be a good data set. There are many possible dependent variables – prices of various fast-food items. The key variable is the fraction of the population that is black, along with controls for poverty, income, housing values, and so on. These data were also used in a famous study by David Card and Alan Krueger on estimation of minimum wage effects on employment. See the book by Card and Krueger, *Myth and Measurement*, 1997, Princeton University Press.

**EARNS.RAW**

**Source:** *Economic Report of the President*, 1989, Table B-47. The data are for the nonfarm business sector.

**Used in Text:** pages 344-345, 379-380, 387

**Notes:** These data could be usefully updated, but changes in reporting conventions in more recent *ERP*s may make that difficult.

**ENGIN.RAW**

**Source:** Thada Chaisawangwong, a graduate student at MSU, used these data for a term project. They come from the Material Requirement Planning Survey carried out in Thailand during 1998.

**Used in Text:** not used

**Notes:** This is a nice change-of-pace from wage data sets for the United States. These data are for engineers in Thailand, and should represent a more homogeneous group than data sets that consist of people in a variety of occupations. Plus, the starting salary is also provided in the data set, so factors affecting wage growth can be studied in addition to factors affecting the wage at any point in time.

**EZANDERS.RAW**

**Source:** L.E. Papke (1994), "Tax Policy and Urban Development: Evidence from the Indiana Enterprise Zone Program," *Journal of Public Economics* 54, 37-49.

Professor Papke kindly provided these data.

**Used in Text:** pages 357-358

**Notes:** These are actually monthly unemployment claims for the Anderson enterprise zone. Papke used annualized data, across many zones and non-zones, in her analysis.

**EZUNEM.RAW**

**Source:** See EZANDERS.RAW

**Used in Text:** pages 450-451, 479

**Notes:** A very good project is to have students analyze enterprise, or empowerment, zone policies in their home states. Many states now have such programs. A few years of panel data over, say, cities or zip codes, could make a nice study.

**FAIR.RAW**

**Source:** R.C. Fair (1996), "Econometrics and Presidential Elections," *Journal of Economic Perspectives* 10, 89-102.

The data set is reported in the article.

**Used in Text:** pages 343-344, 420, 421-422

**Notes:** This is a data set begging to be updated. One could check Professor Fair's web site at Yale University. In any case, he spells out the definitions carefully in the above article. Students might want to try their own hands at predicting the most recent election outcome, but they should be restricted to no more than a handful of explanatory variables.

**FERTIL1.RAW**

**Source:** W. Sander, "The Effect of Women's Schooling on Fertility," *Economics Letters* 40, 229-233.

Professor Sander kindly provided the data, which are a subset of what he used in his article. He compiled the data from various years of the National Opinion Resource Center's General Social Survey.

**Used in Text:** pages 427-429, 454, 512, 596, 651

**Notes:** It would be very interesting to analyze a similar data set for a developing country, especially where efforts have been made to emphasize birth control.

**FERTIL2.RAW**

**Source:** These data were obtained by James Heakins, a former MSU undergraduate, for a term project. They come from Botswana's 1988 Demographic and Health Survey.

**Used in Text:** pages 518-519

**Notes:** Currently, this data set is used only in one computer exercise. Since the dependent variable of interest – number of living children or number of children every born – is a count variable, the Poisson regression model can be used. However, it requires some care to combine Poisson regression with an endogenous explanatory variable (*educ*). I refer you to Chapter 19 of my book *Econometric Analysis of Cross Section and Panel Data*. Even in the context of linear models, much can be done beyond Computer Exercise 15.13. At a minimum, the binary indicators for various religions can be added. You could also interact *educ* with some of the other exogenous explanatory variables.

**FERTIL3.RAW**

**Source:** L.A. Whittington, J. Alm, and H.E. Peters (1990), "Fertility and the Personal Exemption: Implicit Pronatalist Policy in the United States," *American Economic Review* 80, 545-556.

The data are given in the article.

**Used in Text:** 338-340, 349, 357, 358, 378-379, 382, 388, 421, 618, 636, 642, 643

**FISH.RAW**

**Source:** K Graddy (1995), "Testing for Imperfect Competition at the Fulton Fish Market," *RAND Journal of Economics* 26, 75-92.

Professor Graddy's collaborator on a later paper, Professor Joshua Angrist at MIT, kindly provided me with these data.

**Used in Text:** pages 422-423, 552

**Notes:** This is a very nice example of how one can find exogenous variables to use as instrumental variables. Often, weather conditions can be assumed to affect supply while having a negligible effect on demand. Then, the weather variables are valid instrumental variable candidates for price in the demand equation.

**FRINGE.RAW**

**Source:** F. Vella (1993), "A Simple Estimator for Simultaneous Models with Censored Endogenous Regressors," *International Economic Review* 34, 441-457.

Professor Vella kindly provided the data.

**Used in Text:** pages 595-596

**Notes:** Currently, this data set is used in only one Computer Exercise, to illustrate the Tobit model. But it can be used much earlier. First, one could just ignore the pileup at zero and use a linear model where any of the hourly benefit measures is the dependent variable. Another possibility is to use this data set for a problem set in Chapter 4, after students have read Example 4.10. That example, which uses teacher salary/benefit data at the school level, finds the expected tradeoff, although it appears to less than one-to-one. By contrast, if you do a similar analysis with FRINGE.RAW, you will not find a tradeoff. A positive coefficient on the benefit/salary ratio is not too surprising because we probably cannot control for enough factors, especially when looking across different occupations. The Michigan school-level data is more aggregated than one would like, but it does restrict attention to a more homogeneous group: high school teachers in Michigan.

**GPA1.RAW**

**Source:** Christopher Lemmon, a former MSU undergraduate, collected these data from surveying other MSU students in Fall 1994.

**Used in Text:** pages 75-76, 82, 128, 159-160, 223-224, 252, 282-283, 287, 803

**Notes:** This is a nice example of how students can obtain an original data set by focusing locally, and carefully composing a survey.

**GPA2.RAW**

**Source:** For confidentiality reasons, I cannot divulge the source of these data. I can say that they come from a fairly large research university that also supports men's and women's athletics at the Division I level.

**Used in Text:** pages 106, 180, 204, 206, 214, 250, 253

**GPA3.RAW**

**Source:** See GPA2.RAW

**Used in Text:** pages 238-240, 262, 284, 456

**HPRICE1.RAW**

**Source:** Collected from the real estate pages of the *Boston Globe* during 1990. These are homes selling in the Boston, MA area.

**Used in Text:** pages 110, 154, 160, 164, 207, 215, 216, 225, 267, 269, 285, 293

**Notes:** Typically, it is very easy to obtain data on selling prices and characteristics of homes, using publicly-available data bases. It is interesting to match the information on houses with other information – such as local crime rates, quality of the local schools, pollution levels, and so on – and estimate the effects of such variables on housing prices.

**HPRICE2.RAW**

**Source:** D. Harrison and D.L. Rubinfeld (1978), "Hedonic Housing Prices and the Demand for Clean Air," by Harrison, D. and D.L.Rubinfeld, *Journal of Environmental Economics and Management* 5, 81-102.

Diego Garcia, a former Ph.D. student in economics at MIT, kindly provided these data, which he obtained from the book *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, by D.A. Belsey, E. Kuh, and R. Welsch, 1990. New York: Wiley.

**Used in Text:** pages 108, 131, 186-187, 192-194

**Notes:** The census contains rich information on things like median housing prices, median income levels, average family size, and so on for fairly small geographical areas. If such data can be merged with pollution data, one can update the Harrison and Rubinfeld study. Presumably, this has been done often in academic journals.

**HPRICE3.RAW**

**Source:** K.A. Kiel and K.T. McClain (1995), "House Prices During Siting Decision Stages: The Case of an Incinerator from Rumor Through Operation," *Journal of Environmental Economics and Management* 28, 241-255.

Professor McClain kindly provided me with the data, of which I used only a subset.

**Used in Text:** not used

**Notes:** This is a redundant data set; it is essentially the same as KIELMC.RAW

**HSEINV.RAW**

**Source:** D. McFadden (1994), "Demographics, the Housing Market, and the Welfare of the Elderly," in D.A. Wise (ed.), *Studies in the Economics of Aging*. Chicago: University of Chicago Press, 225-285.

The data are contained in the article.

**Used in Text:** pages 348-349, 352, 387, 606-607, 642, 814

**HTV.RAW**

**Source:** J.J. Heckman, J.L. Tobias, and E. Vytlacil, "Simple Estimators for Treatment Parameters in a Latent Variable Framework with an Application to Estimating the Returns to Schooling," National Bureau of Economic Research Working Paper No. 7950, October 2000.

Professor Tobias kindly provided the data, which were obtained from the 1991 National Longitudinal Survey of Youth. For confidentiality reasons, I have included only a subset of the variables.

**Used in Text:** page 522

**Notes:** Because an ability measure is included in the data set, it can be used as another illustration of including proxy variables in regression models. See Chapter 9.

**INFMRT.RAW**

**Source:** *Statistical Abstract of the United States*, 1990 and 1994. (For example, the infant mortality rates come from Table 113 in 1990 and Table 123 in 1994.)

**Used in Text:** pages 314-315, 320

**Notes:** An interesting exercise is to add the percent of the population on AFDC (*afdcper*) to the infant mortality equation. Pooled OLS and first differencing can give very different estimates.

**INJURY.RAW**

**Source:** B.D. Meyer, W.K. Viscusi, and D.L. Durbin (1995), "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment," *American Economic Review* 85, 322-340.

Professor Meyer kindly provided the data.

**Used in Text:** pages 437, 455

**Notes:** This data set can be used to illustrate the Chow test in Chapter 7. In particular, students can test whether the regression functions differ between Kentucky and Michigan. Or, allowing an intercept difference, do the slopes differ?

**INTDEF.RAW**

**Source:** From the *Economic Report of the President*, 1997, Tables B-71 and B-78.

**Used in Text:** pages 357, 359, 410, 519-520

**INTQRT.RAW**

**Source:** From Salomon Brothers, *Analytical Record of Yields and Yield Spreads*, 1990. The people at Salomon Brothers kindly provided the *Record* at no charge when I was an assistant professor at MIT.

**Used in Text:** pages 385-386, 609-610, 616, 620, 621-622, 642, 644

**Notes:** The nice thing about the Salomon Brothers data is that the interest rates are not averaged over a month or quarter – they are end-of-month or end-of-quarter rates. Asset pricing theories apply to such "point-sampled" data, and not to averages over a period. Most other sources report monthly or quarterly averages. This is a good data set to update and test whether current data are more or less supportive of basic asset pricing theories.

**INVEN.RAW**

**Source:** *Economic Report of the President*, 1997, Tables B-4, B-20, B-61, and B-71.

**Used in Text:** pages 388-389, 612-613, 814

**JTRAIN.RAW**

**Source:** H. Holzer, R. Block, M. Cheatham, and J. Knott (1993), "Are Training Subsidies Effective? The Michigan Experience," *Industrial and Labor Relations Review* 46, 625-636.

The authors kindly provided the data.

**Used in Text:** pages 135-136, 160-161, 224-225, 246, 320, 445-446, 457, 463-464, 469, 478, 512-513, 752-754, 764-765, 767, 810

**JTRAIN2.RAW**

**Source:** R.J. Lalonde (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review* 76, 604-620.

Professor Jeff Biddle, at MSU, kindly passed the data along to me. He obtained it from Professor Lalonde.

**Used in Text:** page 597

**Notes:** Professor Lalonde obtained the data from the National Supported Work Demonstration job-training program conducted by the Manpower Demonstration Research Corporation in the mid 1970s. Training status was randomly assigned, so this is essentially experimental data. Computer Exercise 17.15 only looks at the effects of training on subsequent unemployment probabilities. The data set also contains real earnings. Because of random assignment, a simple comparison of means suffices. Nevertheless, a good exercise would be to have the students estimate a Tobit of *re78* on *train*, and obtain estimates of the expected values for those with and without training. These can be compared with the sample average.

**KIELMC.RAW**

**Source:** K.A. Kiel and K.T. McClain (1995), "House Prices During Siting Decision Stages: The Case of an Incinerator from Rumor Through Operation," *Journal of Environmental Economics and Management* 28, 241-255.

Professor McClain kindly provided the data, of which I used only a subset.

**Used in Text:** pages 213-214, 432-436, 453, 455

**LAWSCH85.RAW**

**Source:** Collected by Kelly Barnett for use in a term project. The data come from two sources: *The Official Guide to U.S. Law Schools*, 1986, Law School Admission Services, and *The Gourman Report: A Ranking of Graduate and Professional Programs in American and International Universities*, 1995, Washington, D.C.

**Used in Text:** pages 107, 164, 231

**Notes:** More recent versions of both cited documents are available. One could try a similar analysis for, say, Ph.D. programs in economics. Quality of placements may be a good dependent variable, and measures of graduate program quality could be included among the explanatory variables. Of course, one would want to control for factors describing the incoming class so as to isolate the effect of the program itself.

**LOANAPP.RAW**

**Source:** W.C. Hunter and M.B. Walker (1996), "The Cultural Affinity Hypothesis and Mortgage Lending Decisions," *Journal of Real Estate Finance and Economics* 13, 57-70.

Professor Walker kindly provided the data.

**Used in Text:** 254-255, 287, 321, 595

**Notes:** These data were originally used in a famous study by researchers at the Boston Federal Reserve Bank. See A. Munnell, G.M.B. Tootell, L.E. Browne, and J. McEneaney (1996), "Mortgage Lending in Boston: Interpreting HMDA Data," *American Economic Review* 86, 25-53.

**LOWBRTH.RAW**

**Source: Source:** *Statistical Abstract of the United States*, 1990, 1993, and 1994.

**Used in Text:** not used

**Notes:** This data set can be used very much like INFMRT.RAW. It contains two years of state-level panel data. In fact, it is a superset of INFMRT.RAW. The key is that it contains information on low birth weights, as well as infant mortality. It also contains state identifies, so that several years of more recent data could be added for a term project. Putting in the variable *afcdprc* and its square lead to some interesting findings for pooled OLS and fixed effects (first differencing). After differencing, you can even try using the change in the AFDC payments variable as an instrumental variable for the change in *afdcprc*.

**MATHPNL.RAW**

**Source:** Leslie E. Papke, an economist at MSU, collected these data from the web site that contains information on the Michigan Educational Assessment Program (MEAP):

www.meritaward.state.mi.us. These are district-level data, which Professor Papke kindly provided. She has used building-level data in "The Effects of Spending on Test Pass Rates: Evidence from Michigan," 2001, MSU Department of Economics Working paper.

**Used in Text:** pages 457-458, 480

**MEAP93.RAW**

**Source:** I collected these data from the old Michigan Department of Education web site. (MEAP is now under the auspices of the Michigan Department of Treasury.) See MATHPNL.RAW for the current web site. I used data on most high schools in the state of Michigan for 1993. I dropped some high schools that had suspicious looking data.

**Used in Text:** pages 52-53, 124-126, 155-156, 212, 318-319, 321, 668

**Notes:** Many states have data, at either the district or building level, on student performance and spending. The Michigan site now has data at the student level (where the students are anonymous, of course).

**MLB1.RAW**

**Source:** Collected by George Holmes, a former MSU undergraduate, for a term project. The salary data were obtained from the New York *Times*, April 11, 1993. The baseball statistics are from *The Baseball Encyclopedia*, 9th edition, and the city population figures are from the *Statistical Abstract of the United States*.

**Used in Text:** pages 144-148, 164, 236, 253

**Notes:** The baseball statistics are career statistics through the 1992 season. Players whose race or ethnicity could not be easily determined were not included. It should not be too difficult to obtain the city population and racial composition numbers for Montreal and Toronto for 1993.

**MROZ.RAW**

**Source:** T.A. Mroz (1987), "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica* 55, 765-799.

Professor Ernst R. Berndt, of MIT, kindly provided the data, which he obtained from Professor Mroz.

**Used in Text:** pages 241-243, 251-252, 280-281, 490-491, 501, 506, 508, 536-537, 550
561-563, 569-572, 590-591, 596-597

**MURDER.RAW**

**Source:** From the *Statistical Abstract of the United States*, 1995 (Tables 310 and 357), 1992 (Table 289). The execution data originally come from the U.S. Bureau of Justice Statistics, *Capital Punishment Annual*.

**Used in Text:** pages 479-480, 520

**Notes:** Prosecutors in different counties might pursue the death penalty with different intensities, so it might make sense to collect murder and execution data at the county level. This could be combined with better demographic information at the county level, along with better economic data (say, on wages for various kinds of employment).

**NBASAL.RAW**

**Source:** Collected by Christopher Torrente, a former MSU undergraduate, for a term project. He obtained the salary data and the career statistics from *The Complete Handbook of Pro Basketball*, 1995, Edited by Zander Hollander. New York: Signet. The demographic information (marital status, number of children, and so on) was obtained from the teams' 1994-1995 media guides.

**Used in Text:** pages 216, 255.

**Notes:** A panel version of this data set could be useful for further isolating productivity effects of marital status. One would need to obtain information on enough different players in at least two years, where some players who were not married in the initial year are married in later years. Fixed effects (or first differencing, for two years) is the natural estimation method.

**NYSE.RAW**

**Source:** These are Wednesday closing prices of value-weighted NYSE average, available in many publications. I do not recall the particular source I used. Probably the easiest way to get similar data is to go to the NYSE web site, http://www.nyse.com/.

**Used in Text:** pages 369-370, 387-388, 415, 417, 420-421, 634

**OPENNESS.RAW**

**Source:** D. Romer (1993), "Openness and Inflation: Theory and Evidence," *Quarterly Journal of Economics* 108, 869-903.

The data are included in the article.

**Used in Text:** pages 538, 550-551

**PENSION.RAW**

**Source:** L.E. Papke (2002), "Individual Financial Decisions in Retirement Saving: The Role of Participant-Direction," forthcoming, *Journal of Public Economics.*

Professor Papke kindly provided the data. She collected them from the National Longitudinal Survey of Mature Women, 1991.

**Used in Text:** pages 480-481

**PHILLIPS.RAW**

**Source:** *Economic Report of the President*, 1998, Tables B-42 and B-63.

**Used in Text:** pages 336, 371-372, 388, 389, 390, 397, 408, 521, 611, 626, 629, 632, 642-643, 807

**Notes:** Clearly a good data set to update. Some of the exercises ask the students to do so, to see how the different Phillips curves change when estimated using recent data.

**PNTSPRD.RAW**

**Source:** Collected by Scott Resnick, a former MSU undergraduate, from various newspaper sources.

**Used in Text:** pages 286, 595, 668

**Notes:** The data are for the 1994-1995 men's college basketball seasons. The spread is for the day before the game was played.

**PRISON.RAW**

**Source:** S.D. Levitt (1996), "The Effect of Prison Population Size on Crime Rates: Evidence from Prison Overcrowding Legislation," *Quarterly Journal of Economics* 111, 319-351.

Professor Levitt kindly provided me with the data, of which I used a subset.

**Used in Text:** page 545

**PRMINWGE.RAW**

**Source:** A.J. Castillo-Freeman and R.B. Freeman (1992), "When the Minimum Wage Really Bites: The Effect of the U.S.-Level Minimum Wage on Puerto Rico," in *Immigration and the Work Force*, edited by G.J. Borjas and R.B. Freeman, 177-211. Chicago: University of Chicago Press.

The data are reported in the article.

**Used in Text:** pages 337, 351, 400, 413

**Notes:** Given the ongoing debate on the employment effects of the minimum wage, this would be a great data set to try to update. The coverage rates are the most difficult variables to construct.

**RECID.RAW**

**Source:** C.-F. Chung, P. Schmidt, and A.D. Witte (1991), "Survival Analysis: A Survey," *Journal of Quantitative Criminology* 7, 59-98.

Professor Chung kindly provided the data.

**Used in Text:** pages 580-582, 596

**RDCHEM.RAW**

**Source:** From *Businessweek* R&D Scoreboard, October 25, 1991.

**Used in Text:** pages 66, 159, 199, 211-212, 312-314, 320-321

**Notes:** It would be interesting to collect more recent data and see whether the R&D/firm size relationship has changed.

**RDTELEC.RAW**

**Source:** See RDCHEM.RAW

**Used in Text:** not used

**Notes:** According to these data, the R&D/firm size relationship is different in the telecommunications industry than in the chemical industry: there is pretty strong evidence that R&D intensity decreases with firm size in telecommunications. Of course, that was in 1991. The data could easily be updated, and a panel data set could be constructed.

**RENTAL.RAW**

**Source:** David Harvey, a former MSU undergraduate, collected the data for 64 "college towns" from the 1980 and 1990 United States censuses.

**Used in Text:** page 159, 456, 478

**Notes:** These data can be used in a somewhat crude simultaneous equations analysis, either focusing on one year or pooling the two years. (In the latter case, in an advanced class, you

might have students compute the standard errors robust to serial correlation across the two time periods.) The demand equation would have *ltothsg* as a function of *lrent*, *lavginc*, and *lpop*. The suppy equation would have *ltothsg* as a function of *lrent*, *pctst*, and *lpop.* Thus, in estimating the demand function, *pctstu* is used as an IV for *lrent*. Clearly one can quibble with excluding *pctstu* from the demand equation, but the estimated demand function gives a negative price effect.

Getting information for 2000, and adding many more college towns, would make for a much better analysis. Information on number of spaces in on-campus dormitories would be a big improvement, too.

**RETURN.RAW**

**Source:** Collected by Stephanie Balys, a former MSU undergraduate, from the New York Stock Exchange and *Compustat*.

**Used in Text:** 162

**Notes:** More can be done with this data set. Recently, I discovered that *lsp90* does appear to predict *return* (and the log of the 1990 stock price works better than *sp90*). I am a little suspicious, but you could use the negative coefficient on *lsp90* to illustrate "reversion to the mean."

**SAVING.RAW**

**Source:** Unknown

**Used in Text:** 273-274

**Notes:** I remember entering this data set in the late 1980s, and I am pretty sure it came directly from an introductory econometrics text. But so far my search has been fruitless. If anyone runs across this data set, I would appreciate knowing about it.

**SLEEP75.RAW**

**Source:** J.E. Biddle and D.S. Hamermesh (1990), "Sleep and the Allocation of Time," *Journal of Political Economy* 98, 922-943.

Professor Biddle kindly provided the data.

**Used in Text:** pages 65, 106-107, 161-162, 249, 254, 285

**Notes:** In their article, Biddle and Hamermesh include an hourly wage measure in the sleep equation. An econometric problem is that the hourly wage is missing for those who do not work. Plus, the wage offer may be endogenous (even if it were always observed). Biddle and Hamermesh employ extensions of the sample selection methods in Section 17.5. See their article for details.

**SLP75_81.RAW**

**Source:** See SLEEP75.RAW

**Used in Text:** pages 442-443

**SMOKE.RAW**

**Source:** J. Mullahy (1997), "Instrumental-Variable Estimation of Count Data Models: Applications to Models of Cigarette Smoking Behavior," *Review of Economics and Statistics* 79, 596-593.

Professor Mullahy kindly provided the data.

**Used in Text:** page 180, 278-279, 284-285, 287, 550, 598

**Notes:** If you want to do a "fancy" IV version of Computer Exercise 16.9, you could estimate a reduced form count model for *cigs* using the Poisson regression methods in Section 17.3, and then use the fitted values as an IV for *cigs*. Presumably, this would be for a fairly advanced class.

**TRAFFIC1.RAW**

**Source:** I collected these data from two sources, the 1992 *Statistical Abstract of the United States* (Tables 1009, 1012) and *A Digest of State Alcohol-Highway Safety Related Legislation*, 1985 and 1990, published by the U.S. National Highway Traffic Safety Administration.

**Used in Text:** pages 446-447, 667

**Notes:** In addition to adding recent years, this data set could really use state-level tax rates on alcohol.

**TRAFFIC2.RAW**

**Source:** P.S. McCarthy (1994), "Relaxed Speed Limits and Highway Safety: New Evidence from California," *Economics Letters* 46, 173-179.

Professor McCarthy kindly provided the data.

**Used in Text:** pages 359, 389, 422, 645, 667

**Notes:** Many states have changed maximum speed limits and imposed seat belt laws over the past 25 years. Data similar to those in TRAFFIC2.RAW should be fairly easy to obtain for a particular state.

**TWOYEAR.RAW**

**Source:** T.J. Kane and C.E. Rouse (1995), "Labor-Market Returns to Two- and Four-Year Colleges," *American Economic Review* 85, 600-614.

With Professor Rouse's kind assistance, I obtained the data from her web site at Princeton University.

**Used in Text:** pages 139-142, 165, 321

**Notes:** As possible extensions, students can explore whether the returns to two-year or four-year colleges depend on race or gender. Also, should experience appear as a quadratic?

**VOLAT.RAW**

**Source:** J.D. Hamilton and L. Gang (1996), "Stock Market Volatility and the Business Cycle," *Journal of Applied Econometrics* 11, 573-593.

I obtained these data from the *Journal of Applied Econometrics* data archive at http://qed.econ.queensu.ca/jae/.

**Used in Text:** pages 358-359, 640-641, 642, 644-645

**VOTE1.RAW**

**Source:** From M. Barone and G. Ujifusa, *The Almanac of American Politics*, 1992. Washington, DC: National Journal.

**Used in Text:** pages 35, 41, 163-164, 215, 286, 671

**VOTE2.RAW**

**Source:** See VOTE1.RAW

**Used in Text:** pages 318, 457, 671

**Notes:** These are panel data collected for the 1988 and 1990 U.S. House of Representative elections.

**WAGE1.RAW**

**Source:** These are data from the 1976 Current Population Survey, collected by Henry Farber when he and I were colleagues at MIT in 1988.

**Used in Text:** pages 7, 38, 76-77, 93, 123-124, 180, 190-192, 214, 222-223, 226-228, 232, 235, 254, 260-261, 311, 648

**WAGE2.RAW**

**Source:** M. Blackburn and D. Neumark (1992), "Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials," *Quarterly Journal of Economics* 107, 1421-1436.

Professor Neumark kindly provided the data, of which I used a subset.

**Used in Text:** pages 65-66, 106, 112, 164-165, 212, 214, 252-253, 297-299, 320, 491, 505 518, 521-522, 648

**WAGEPAN.RAW**

**Source:** F. Vella and M. Verbeek (1998), "Whose Wages Do Unions Raise? A Dynamic Model of Unionism and Wage Rate Determination for Young Men," *Journal of Applied Econometrics* 13, 163-183.

I obtained the data from the *Journal of Applied Econometrics* data archive at http://qed.econ.queensu.ca/jae/.

**Used in Text:** pages 465, 471-473, 479

**WAGEPRC.RAW**

**Source:** *Economic Report of the President*, various years.

**Used in Text:** pages 385, 421, 641-642

**Notes:** These monthly data run from January 1964 through October 1987. The consumer price index averages to 100 in 1967.

**WINE.RAW**

**Source:** These data were reported in a New York *Times* article, December 28, 1994.

**Used in Text:** not used

**Notes:** The dependent variables *deaths*, *heart*, and *liver* can be regressed against *alcohol* as nice simple regression examples.