DATA SET HANDBOOK

Introductory Econometrics: A Modern Approach
Tradução da 6ª edição norte-americana
Jeffrey M. Wooldridge

This document contains a listing of all data sets that are provided with the *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.

Special thanks to Edmund Wooldridge, who updated the page numbers for the sixth edition.

401K

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.

Notes: This data set is used in a variety of ways in the text. One additional possibility is to investigate whether the coefficients from the regression of *prate* on *mrate*, log(*totemp*) differ by whether the plan is a sole plan. The Chow test (see Section 7.4), and the less restrictive version that allows different intercepts, can be used.

401KSUBS

Source: A. Abadie (2003), "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics* 113, 231-263.

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

Notes: This data set can also be used to illustrate the binary response models, probit and logit, in Chapter 17, where, say, *pira* (an indicator for having an individual retirement account) is the dependent variable, and *e401k* [the 401(k) eligibility indicator] is the key explanatory variable.

ADMNREV

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

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 or later years can be added, forming 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

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 Yale University Department of Economics. He originally obtained the data from a survey by *Psychology Today*.

Notes: This is an interesting data set for problem sets starting in Chapter 7. Even though *naffairs* (number of extramarital affairs a woman reports) is a count variable, a linear model can be used to model its conditional mean as an approximation. Or, you could ask the students to estimate a linear probability model for the binary indicator *affair*, equal to one of the woman reports having any extramarital affairs. One possibility is to test whether putting the single marriage rating variable, *ratemarr*, is enough, against the alternative that a full set of dummy variables is needed; see pages 239-240 for a similar example. This is also a good data set to illustrate Poisson regression (using *naffairs*) in Section 17.3 or probit and logit (using *affair*) in Section 17.1.

AIRFARE

Source: Jiyoung Kwon, a former doctoral student 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.

Notes: This data set nicely illustrates the different estimates obtained when applying pooled OLS, random effects, and fixed effects.

ALCOHOL

Source: Terza, J.V. (2002), "Alcohol Abuse and Employment: A Second Look," *Journal of Applied Econometrics* 17, 393-404.

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

APPLE

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, which were obtained from a telephone survey conducted by the Institute for Public Policy and Social Research at MSU.

Notes: This data set is close to a true experimental data set because the price pairs facing a family were randomly determined. In other words, the family head was presented with prices for the eco-labeled and regular apples, and then asked how much of each kind of apple the family would buy at the given prices. As predicted by basic economics, the own price effect is negative (and strong) and the cross price effect is positive (and strong). While the main dependent variable, *ecolbs*, piles up at zero, estimating a linear model is still worthwhile. Interestingly, because the survey design induces a strong positive correlation between the prices of eco-labeled and regular apples, there is an omitted variable problem if either of the price variables is dropped from the demand equation. A good exam question is to show a simple regression of *ecolbs* on *ecoprc* and then a multiple regression on both prices, and ask students to decide whether the price variables must be positively or negatively correlated.

APPROVAL

Source: Harbridge, L., J. Krosnick, and J.M. Wooldridge (forthcoming), "Presidential Approval and Gas Prices: Sociotropic or Pocketbook Influence?" in *New Explorations in Political Psychology*, ed. J. Krosnick. New York: Psychology Press (Taylor and Francis Group)

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

ATHLET1

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.

Notes: These data were collected by Patrick Tulloch, an MSU economics major, 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

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

Notes: These data were collected by Paul Anderson, an MSU economics major, 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. Extended these data to obtain a long stretch of panel data and other "natural" rivals could be very interesting.

ATTEND

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, who completed his Ph.D. in economics at MSU, provided helpful hints.

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

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.

BARIUM

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.

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

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

BEAUTY

Source: Hamermesh, D.S. and J.E. Biddle (1994), "Beauty and the Labor Market," *American Economic Review* 84, 1174-1194.

Professor Hamermesh kindly provided me with the data. For manageability, I have included only a subset of the variables, which results in somewhat larger sample sizes than reported for the regressions in the Hamermesh and Biddle paper.

BIG9SALARY

Source: O. Baser and E. Pema (2003), "The Return of Publications for Economics Faculty," *Economics Bulletin* 1, 1-13.

Professors Baser and Pema kindly provided the data.

Notes: This is an unbalanced panel data set in the sense that as many as three years of data are available for each faculty member but where some have fewer than three years. It is not clear that something like a fixed effects or first differencing analysis makes sense: in effect, approaches that remove the heterogeneity control for too much by controlling for unobserved heterogeneity which, in this case, includes faculty intelligence, talent, and motivation. Presumably these factors enter into the publication index. It is hard to think we want to hold the main factors driving productivity fixed when trying to measure the effect of productivity on salary. Pooled OLS regression with "cluster robust" standard errors seems more natural.

On the other hand, if we want to measure the return to having a degree from a top 20 Ph.D. program then we would want to control for factors that cause selection into a top 20 program. Unfortunately, this variable does not change over time, and so FD and FE are not applicable.

BWGHT

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.

BWGHT2

Source: Dr. Zhehui Luo, a professor of epidemiology and biostatistics at MSU, kindly provided these data. She obtained them from state files linking birth and infant death certificates, and from the National Center for Health Statistics natality and mortality data.

Notes: There are many possibilities 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 Exercise C6.10. 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 might ask students about predicted values less than zero or greater than 10.

CAMPUS

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.

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, policy variables – such as a "safe house" for women on campus, as was started at MSU in 1994 – could be added as explanatory variables. The crime rate in the host town would be a good control.

CARD

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.

Notes: Computer Exercise C15.3 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 correlation between *nearc4* and *IQ*, once the other explanatory variables are netted out, is arguably zero. (At least, it is not statistically different from zero.) In other words, *nearc4* fails the exogeneity requirement in a simple linear model but it passes – at least using the crude test described above – if controls are added to the wage equation.

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

CATHOLIC

Source: Altonji, J.G., T.E. Elder, and C.R. Taber (2005), "An Evaluation of Instrumental Variable Strategies for Estimating the Effects of Catholic Schooling," *Journal of Human Resources* 40, 791-821.

Professor Elder kindly provided a subset of the data, with some variables stripped away for confidentiality reasons.

CEMENT

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.

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 not been seasonally adjusted.

CENSUS2000

Source: Obtained from the United States Census Bureau by Professor Alberto Abadie of the Harvard Kennedy School of Government.

Professor Abadie kindly provided the data.

CEOSAL1

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

Notes: This kind of data collection is relatively easy for students just learning data analysis, and the findings can be interesting. A good term project is to have students collect a similar data set using a more recent issue of *Businessweek*, and to find additional variables that might explain differences in CEO compensation. My impression is that the public is still interested in CEO compensation.

An interesting question is whether the list of explanatory variables included in this data set now explain less of the variation in log(*salary*) than they used to.

CEOSAL2

Source: See CEOSAL1

Notes: Compared with CEOSAL1, in this CEO data set more information about the CEO, rather than about the company, is included.

CHARITY

Source: P.H. Franses and R. Paap (2001), *Quantitative Models in Marketing Research*. Cambridge: Cambridge University Press.

Professor Franses kindly provided the data.

Notes: This data set can be used to illustrate probit and Tobit models, and to study the linear approximations to them.

CONSUMP

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.

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

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.

Notes: You could use these data to illustrate simple regression when the population intercept should be zero: no corn pixels should predict no corn planted. The same can be done with the soybean measures in the data set.

COUNTYMURDERS

Source: Compiled by J. Monroe Gamble for a Summer Research Opportunities Program (SROP) at Michigan State University, Summer 2014. Monroe obtained data from the U.S. Census Bureau, the FBI Uniform Crime Reports, and the Death Penalty Information Center.

CPS78_85

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.

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

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

Notes: This is much bigger than the other CPS data sets even though the sample is restricted to married women. (CPS91 contains many more observations than MROZ, too.) In addition to the usual human capital variables for the women in the sample, we have information on the husband. Therefore, we can estimate a labor supply function as in Chapter 16, although the validity of potential experience as an IV for log(*wage*) is questionable. (MROZ contains an actual experience variable.) Perhaps more convincing is to add *hours* to the wage offer equation, and instrument hours with indicators for young and old children. This data set also contains a union membership indicator.

The web site for the National Bureau of Economic Research makes it very easy now to download CPS data files in a variety of formats. Go to http://www.nber.org/data/cps_basic.html.

CRIME1

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.

CRIME2

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 1985. Unfortunately, I do not have the list of cities.

Notes: Very rich crime data sets, at the county, or even city, level, can be collected using the FBI's *Uniform Crime Reports*. These data can be matched up with demographic and economic

data, at least for census years. The *County and City Data Book* contains a variety of statistics, but the years do not always match up. These data sets can be used investigate issues such as the effects of casinos on city or county crime rates.

CRIME3:

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

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

CRIME4

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.

Notes: Computer Exercise C16.7 shows that variables that might seem to be good instrumental variable candidates are not always so good, especially after applying a transformation such as differencing across time. You could have the students do an IV analysis for just, say, 1987.

DISCRIM

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 set.

Notes: If you want to assign a common final project, this would be a good data set. There are many possible dependent variables, namely, 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, for a detailed analysis.

DRIVING

Source: Freeman, D.G. (2007), "Drunk Driving Legislation and Traffic Fatalities: New Evidence on BAC 08 Laws," *Contemporary Economic Policy* 25, 293--308.

Professor Freeman kindly provided the data.

Notes: Several more years of data are now available and may further shed light on the effectiveness of several traffic laws.

EARNS

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

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

ECONMATH

Source: Compiled by Professor Charles Ballard, Michigan State University Department of Economics.

Professor Ballard kindly provided the data.

ELEM94_95

Source: Culled from a panel data set used by Leslie Papke in her paper "The Effects of Spending on Test Pass Rates: Evidence from Michigan" (2005), *Journal of Public Economics* 89, 821-839.

Notes: Starting in 1995, the Michigan Department of Education stopped reporting average teacher benefits along with average salary. This data set includes both variables, at the school level, and can be used to study the salary-benefits tradeoff, as in Chapter 4. There are a few suspicious benefits/salary ratios, and so this data set makes a good illustration of the impact of outliers in Chapter 9.

EXPENDSHARES

Source: Blundell, R., A. Duncan, and K. Pendakur (1998), "Semiparametric Estimation and Consumer Demand," *Journal of Applied Econometrics* 13, 435-461.

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

Notes: The dependent variables in this data set – the expenditure shares – are necessarily bounded between zero and one. The linear model is at best an approximation, but the usual IV estimator likely gives good estimates of the average partial effects.

ENGIN

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

Notes: This is a nice change of pace from wage data sets for the United States. These data are for engineers in Thailand, and represents a more homogeneous group than data sets that consist of people across a variety of occupations. Plus, the starting salary is also provided in the data set, so factors affecting wage growth – and not just wage levels at a given point in time – can be studied. This is a good data set for a common term project that tests basic understanding of multiple regression and the interpretation of models with a logarithm for a dependent variable.

EZANDERS

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.

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

EZUNEM

Source: See EZANDERS

Notes: A very good project is to have students analyze enterprise, empowerment, or renaissance zone policies in their home states. Many states now have such programs. A few years of panel data straddling periods of zone designation, at the city or zip code level, could make a nice study.

FAIR

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

The data set is provided in the article.

Notes: An updated version of this data set, through the 2004 election, is available at Professor Fair's web site at Yale University: http://fairmodel.econ.yale.edu/rayfair/pdf/2001b.htm. 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 because of the small sample size.

FERTIL1

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.

Notes: (1) Much more recent data can be obtained from the National Opinion Research Center website, http://www.norc.org/GSS+Website/Download/. Very rich pooled cross sections can be constructed to study a variety of issues – not just changes in fertility over time.

(2) It would be interesting to analyze a similar data set for a developing country, especially where efforts have been made to emphasize birth control. Some measure of access to birth control could be useful if it varied by region. Sometimes, one can find policy changes in the advertisement or availability of contraceptives.

FERTIL2

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.

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 discussed in Chapter 17 can be used. However, some care is required 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 C15.2. At a minimum, the binary indicators for various religions can be added as controls. One might also interact the schooling variable, *educ*, with some of the exogenous explanatory variables.

FERTIL3

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.

FISH

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.

Notes: This is a nice example of how to go about finding exogenous variables to use as instrumental variables. Often, weather conditions can be assumed to affect supply while having a negligible effect on demand. If so, the weather variables are valid instrumental variables for price in the demand equation. It is a simple matter to test whether prices vary with weather conditions by estimating the reduced form for price.

FRINGE

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.

Notes: Currently, this data set is used in only one Computer Exercise – to illustrate the Tobit model. 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, 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

Source: Christopher Lemmon, a former MSU undergraduate, collected these data from a survey he took of MSU students in Fall 1994.

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

GPA2

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

GPA3

Source: See GPA2

HPRICE1

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

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

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.

Notes: The census contains rich information on variables such as 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 in academic journals.

HSEINV

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.

HTV

Source: J.J. Heckman, J.L. Tobias, and E. Vytlacil (2003), "Simple Estimators for Treatment Parameters in a Latent-Variable Framework," *Review of Economics and Statistics* 85, 748-755.

Professor Tobias kindly provided the data, which were obtained from the 1991 National Longitudinal Survey of Youth. All people in the sample are males age 26 to 34. For confidentiality reasons, I have included only a subset of the variables used by the authors.

Notes: Because an ability measure is included in this data set, it can be used as another illustration of including proxy variables in regression models. See Chapter 9. Also, one can try the IV procedure with the ability measure included as an exogenous explanatory variable.

INFMRT

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.)

Notes: An interesting exercise is to add the percentage of the population on AFDC (*afdcper*) to the infant mortality equation. Pooled OLS and first differencing can give very different estimates. Adding the years 1998 and 2002 and applying fixed effects seems natural. Intervening years can be added, too, although variation in the key variables from year to year might be minimal.

INJURY

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.

Notes: This data set also 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 for different intercepts for the two states, do the slopes differ? A good lesson from this example is that a small *R*-squared is compatible with the ability to estimate the effects of a policy. Of course, for the Michigan data, which has a smaller sample size, the estimated effect is much less precise (but of virtually identical magnitude).

INTDEF

Source: *Economic Report of the President*, 2004, Tables B–64, B–73, and B–79.

INTQRT

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

Notes: A nice feature of 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

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

JTRAIN

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.

JTRAIN2

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 set along to me. He obtained it from Professor Lalonde.

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 C17.8 looks only at the effects of training on subsequent unemployment probabilities. For illustrating the more advanced methods in Chapter 17, 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 averages.

JTRAIN3

Source: R.H. Dehejia and S. Wahba (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association* 94, 1053-1062.

Professor Sergio Firpo, at the University of British Columbia, has used this data set in his recent work, and he kindly provided it to me. This data set is a subset of that originally used by Lalonde in the study cited for JTRAIN2.

KIELMC

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.

LAWSCH85

Source: Collected by Kelly Barnett, an MSU economics student, 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.

Notes: More recent versions of both cited documents are available. One could try a similar analysis for, say, MBA programs or Ph.D. programs in economics. Quality of placements may be a good dependent variable, and measures of business school or 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

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.

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

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

Notes: This data set can be used very much like INFMRT. It contains two years of state-level panel data. In fact, it is a superset of INFMRT. 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 leads 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

Source: Dr. Leslie Papke, an economics professor at MSU, collected these data from Michigan Department of Education web site, www.michigan.gov/mde. 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" (2005), *Journal of Public Economics* 89, 821-839.

MEAP00

Source: Michigan Department of Education, www.michigan.gov/mde

MEAP01

Source: Michigan Department of Education, www.michigan.gov/mde

Notes: This is another good data set to compare simple and multiple regression estimates. The expenditure variable (in logs, say) and the poverty measure (*lunch*) are negatively correlated in this data set. A simple regression of *math*4 on *lexppp* gives a negative coefficient. Controlling for *lunch* makes the spending coefficient positive and significant.

MEAP93

Source: I collected these data from the old Michigan Department of Education web site. See MATHPNL 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.

Notes: Many states have data, at either the district or building level, on student performance and spending. A good exercise in data collection and cleaning is to have students find such data for a particular state, and to put it into a form that can be used for econometric analysis.

MEAPSINGLE

Source: Collected by Professor Leslie Papke, an economics professor at MSU, from the Michigan Department of Education web site, www.michigan.gov/mde, and the U.S. Census Bureau. Professor Papke kindly provided the data.

MINWAGE

Source: P. Wolfson and D. Belman (2004), "The Minimum Wage: Consequences for Prices and Quantities in Low-Wage Labor Markets," *Journal of Business & Economic Statistics* 22, 296-311.

Professor Belman kindly provided the data.

Notes: The sectors corresponding to the different numbers in the data file are provided in the Wolfson and Bellman and article.

MLB1

Source: Collected by G. Mark 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*.

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. Of course, the data can be pretty easily obtained for more recent players.

MROZ

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.

MURDER

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*.

Notes: The data set COUNTYMURDERS includes information on executions and murder rates at the county level, and provides more variation.

NBASAL

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.

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

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

OKUN

Source: *Economic Report of the President*, 2007, Tables B–4 and B–42.

OPENNESS

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

The data are included in the article.

PENSION

Source: L.E. Papke (2004), "Individual Financial Decisions in Retirement Saving: The Role of Participant-Direction," *Journal of Public Economics* 88, 39-61.

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

PHILLIPS

Source: *Economic Report of the President*, 2004, Tables B–42 and B–64.

PNTSPRD

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

Notes: The data are for the 1994-1995 men's college basketball seasons. The spread is for the day before the game was played. One might collect more recent data and determine whether the spread has become a less accurate predictor of the actual outcome in more recent years. In other words, in the simple regression of the actual score differential on the spread, is the variance larger in more recent years. (We should fully expect the slope coefficient not to be statistically different from one.)

PRISON

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.

PRMINWGE

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.

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

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.

RDCHEM

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

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

RDTELEC

Source: See RDCHEM

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 for more advanced courses.

RENTAL

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

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 supply 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 2010, 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

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

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

Source: Unknown

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

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.

Notes: In their article, Biddle and Hamermesh include an hourly wage measure in the sleep equation. An econometric problem that arises 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

Source: See SLEEP75

SMOKE

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.

Notes: If you want to do a "fancy" IV version of Computer Exercise C16.1, 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

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.

Notes: In addition to adding recent years, this data set could really use state-level tax rates on alcohol. Other important law changes include defining driving under the influence as having a blood alcohol level of .08 or more, which many states have adopted since the 1980s. The trend really picked up in the 1990s and continued through the 2000s. The data set DRIVING is more complete and more recent, but it is also more complicated.

TRAFFIC2

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.

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

state. One should combine this information with changes in a state's blood alcohol limit and the passage of per se and open container laws.

TWOYEAR

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.

Notes: As possible extensions, students can explore whether the returns to two-year or four-year colleges depend on race or gender. This is partly done in Problem 7.9 but where college is aggregated into one number. Also, should experience appear as a quadratic in the wage specification?

VOLAT

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/

VOTE1

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

VOTE2

Source: See VOTE1

Notes: These are panel data, at the Congressional district level, collected for the 1988 and 1990 U.S. House of Representative elections. Of course, much more recent data are available, possibly even in electronic form.

VOUCHER

Source: Rouse, C.E. (1998), "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program," *Quarterly Journal of Economics* 113, 553-602.

Professor Rouse kindly provided the original data set from her paper.

Notes: This is a condensed version of the data set used by Professor Rouse. The original data set had missing information on many variables, including pre-program and post-program test scores. I did not impute any missing data and have dropped observations that were unusable without filling in missing data. There are 990 students in the current data set but pre-program test scores are available for only 328 of them.

This is a good example of where eligibility for a program is randomized but participation need not be. In addition, even if we look at just the effect of eligibility (captured in the variable *selectyrs*) on the math test score (*mnce*), we need to confront the fact that attrition (students leaving the district) can bias the results. Controlling for the pre-policy test score, *mnce90*, can help – but at the cost of losing two-thirds of the observations. A simple regression of *mnce* on *selectyrs* followed by a multiple regression that adds *mnce90* as a control is informative.

The *selectyrs* dummy variables can be used as instrumental variables for the *choiceyrs* variable to try to estimate the effect of actually participating in the program (rather than estimating the so-called *intention-to-treat* effect). Computer Exercise C15.11 steps through the details.

WAGE1

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

Notes: Barry Murphy, of the University of Portsmouth in the UK, has pointed out that for several observations the values for *exper* and *tenure* are in logical conflict. In particular, for some workers the number of years with current employer (*tenure*) is greater than overall work experience (*exper*). At least some of these conflicts are due to the definition of *exper* as "potential" work experience, but probably not all. Nevertheless, I am using the data set as it was supplied to me.

WAGE2

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 just the data for 1980.

Notes: As with WAGE1, there are some clear inconsistencies among the variables *tenure*, *exper*, and *age*. I have not been able to track down the source of the inconsistency, and so any changes would be effectively arbitrary. Instead, I am using the data as provided by the authors of the above *QJE* article.

WAGEPAN

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/. The JAE data archive is generally a nice resource for undergraduates looking to replicate or extend a published study.

WAGEPRC

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

Notes: These monthly data run from January 1964 through October 1987. The consumer price index averages to 100 in 1967. An updated set of data can be obtained electronically from http://www.gpo.gov/fdsys/browse/collection.action?collectionCode=ERP.

WINE

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

Notes: The dependent variables *deaths*, *heart*, and *liver* each can be regressed on *alcohol* as nice simple regression examples. The conventional wisdom is that wine is good for the heart but not for the liver, something that is apparent in the regressions. Because the number of observations is small, this can be a good data set to illustrate calculation of the OLS estimates and statistics.