Supplementary material: Simple Regression Models

Jan Malte Lichtenberg lichtenberg@mpib-berlin.mpg.de
Özgür Şimşek ozgur@mpib-berlin.mpg.de
Center for Adaptive Behavior and Cognition
Max Planck Institute for Human Development
Lentzeallee 94, 14195 Berlin, Germany

Editor: Tatiana V. Guy, Miroslav Kárný, David Rios-Insua, David H. Wolpert

Abstract

This supplementary material contains learning curves for individual data sets that have not been presented in the main article. It also contains detailed descriptions and source descriptions of all used data sets.

Appendix A. Figures

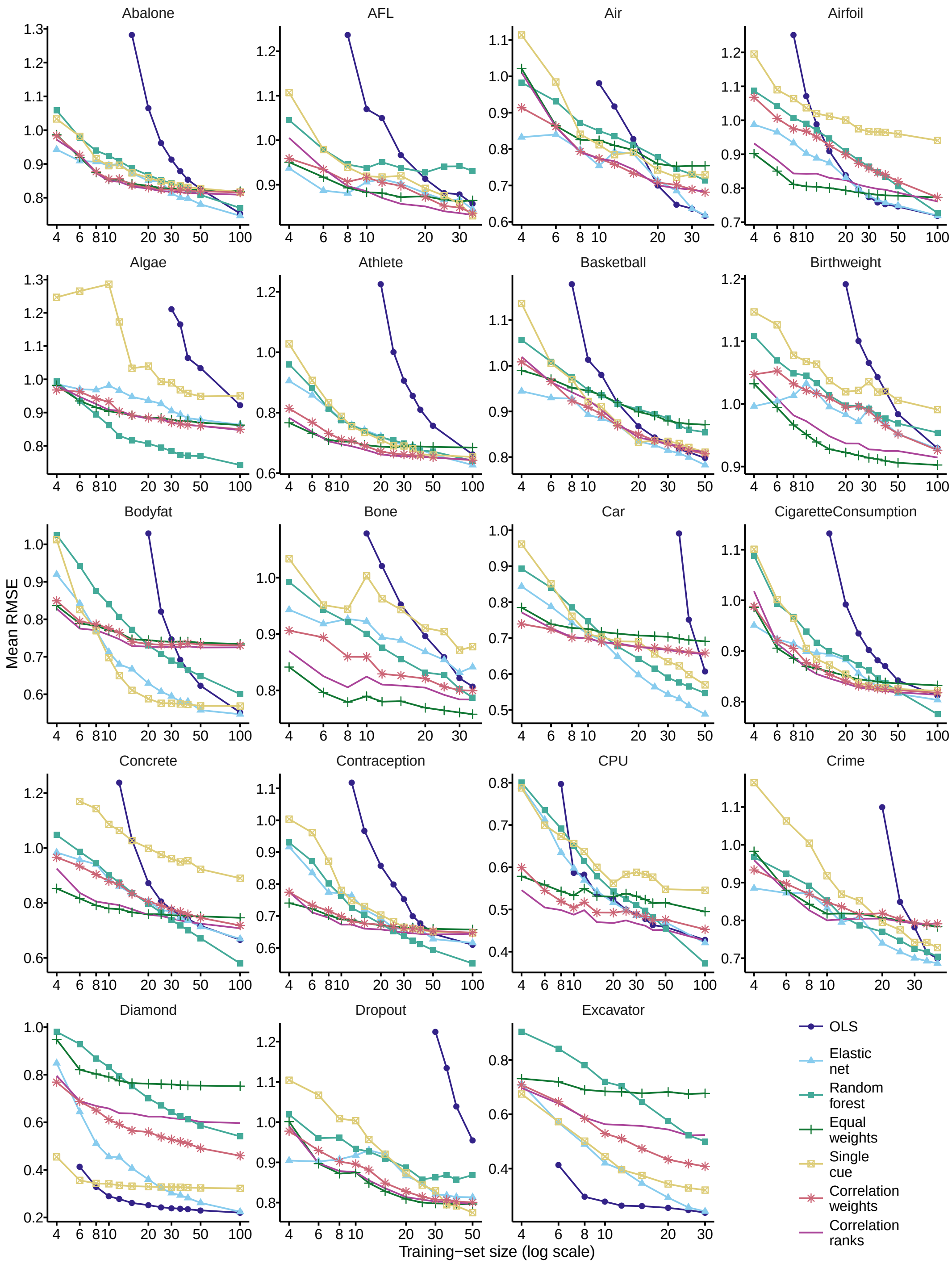

Figure A.1: Learning curves. Data sets 1 to 18. OLS = ordinary least squares.

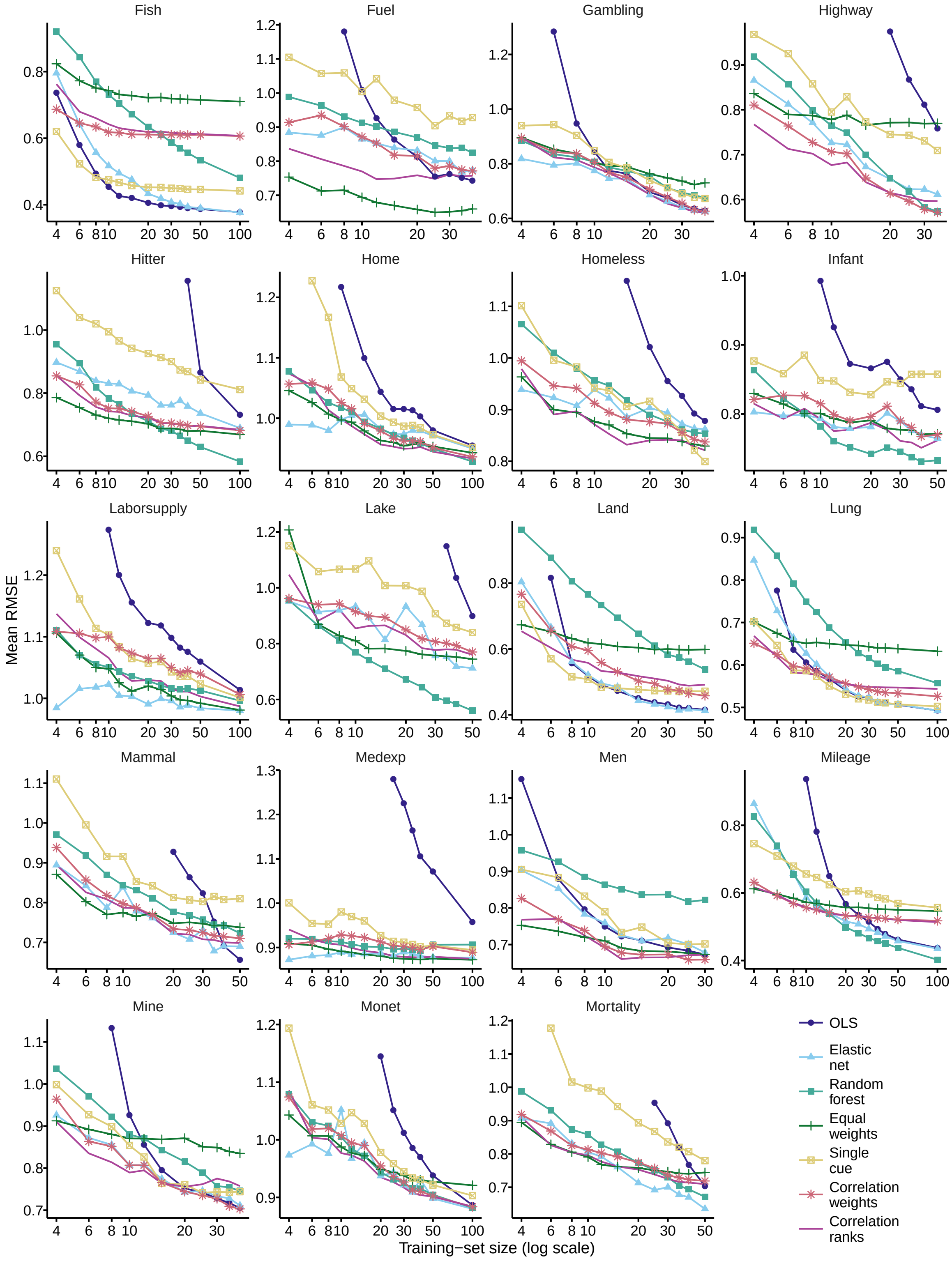

Figure A.2: Learning curves. Data sets 19 to 37. OLS = ordinary least squares.

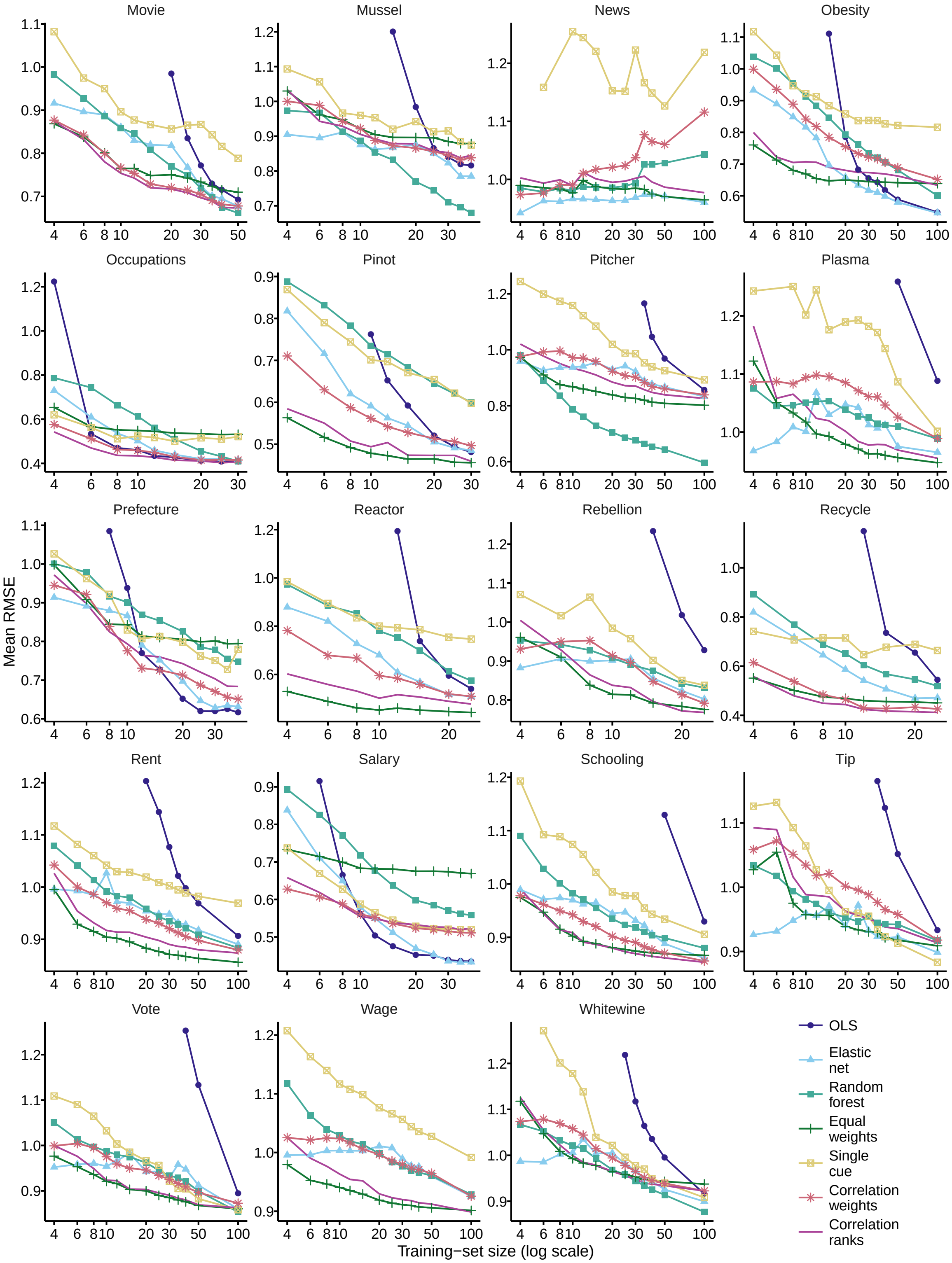

Figure A.3: Learning curves. Data sets 38 to 57. OLS = ordinary least squares.

Appendix B. Data sets

Abalone Objects: 4177 abalones (sea snails). Criterion: age (measured in visible rings). Attributes: sex, length, diameter, height, whole weight, shucked weight, viscera weight, shell weight. Source: This data set comes from a study by Nash et al. (1994). It is available from the UCI Machine Learning Repository (Bache and Lichman, 2013).

AFL Objects: 41 Australian Football League (AFL) games at the Melbourne Cricket Ground in 1993 and 1994. Criterion: attendance. Attributes: forecasted maximum temperature on the day of the game, total attendance at other AFL games in Melbourne and Geelong on the day of the game, total membership in the two clubs whose teams were playing, number of players in the top 50 who participated in the game, number of days since the earliest game of the season. Source: This data set was assembled by Rowan Todd and Mark McNaughton for a class project at the University of Queensland in a statistics course taught by Margaret Mackisack. The data sources were *The Football Bible '94* by Rex Hunt, *The Weekend Australian*, *Inside Football*, and *Football Record*. The data set is available from OzDASL data library (Smyth, 2011), where it is listed with the name *AFL Crowd Attendance at the MCG*.

Air Objects: 41 cities in the United States. Criterion: annual mean concentration of sulfur dioxide. Attributes: average annual temperature, number of manufacturing enterprises employing 20 or more workers, population, average annual wind speed, average annual rainfall, average number of days with rainfall per year. Source: The data were gathered by Sokal and Rohlf (1981) from several publications of the United States government. The data set is reported in a book by Hand et al. (1994) with identifying number 26 and label *air pollution in US cities*.

Airfoil Objects: 1503 airfoils at various wind tunnel speeds and angles of attack. Criterion: scaled sound pressure level, in decibels. Attributes: frequency, angle of attack, chord length, free-stream velocity, suction side displacement thickness. Source: This data set comes from a study by Brooks et al. (1989). It is available from the UCI Machine Learning Repository (Bache and Lichman, 2013).

Algae Objects: 340 samples from European rivers taken over a period of approximately one year. Criterion: density of algae type a. Attributes: concentrations of eight chemicals, season (fall, winter, spring, summer), river size (small, medium, large), fluid velocity (low, medium, high). Source: The data set is from the 1999 Computational Intelligence and Learning (COIL) competition. It is available from the UCI data repository (Bache and Lichman, 2013), where it is labeled *COIL 1999 competition data*.

Athlete Objects: 202 nationally-ranked athletes in Australia. Criterion: blood hemoglobin concentration. Attributes: body mass index, sum of skin folds, percent body fat, lean body mass, height, weight, sex, the sport the athlete competes in (basketball, field, gymnastics, netball, rowing, track 400m, swimming, sprint, tennis, water polo). Source: The data were collected by Telford and Cunningham (1991) at the Australian Institute of Sport. The data set is reported by Maindonald and Braun (2010) and is available from associated R package *DAAG* (Maindonald and Braun, 2013) with label *ais*.

Basketball Objects: 96 basketball players. Criterion: points scored per minute. Attributes: assists per minute, height, time played, age. Source: The data set is reported by Simonoff (1996) and is available from a website maintained by the author (Simonoff, 2015), where it is labeled *baskball.dat*.

Birthweight Objects: 189 newborns. Criterion: birth weight. Attributes: age of mother, weight of mother at last menstrual period, race (white, black, other), number of previous premature labors, number of physician visits during the first trimester, presence of uterine irritability, whether the mother smoked during pregnancy, whether the mother has a history of hypertension. Source: The data were collected at Baystate Medical Center in Springfield, Massachusetts in 1986 (Hosmer and Lemeshow, 2000). The data set is electronically available from R package *MASS* (Venables and Ripley, 2002; Ripley et al., 2013), where it is labeled *birthwt*.

Bodyfat Objects: 252 males. Criterion: percentage of body fat determined by underwater weighing. Attributes: age, weight, height, and various body circumference measurements: neck, chest, abdomen, hip, thigh, knee, ankle, biceps, forearm, wrist. Source: The data were collected by Penrose et al. (1985). The data set is available from StatLib (StatLib: Data, software and news from the statistics community) with label *bodyfat*.

Bone Objects: 42 male skeletons buried in coffins. Criterion: nitrogen content. Attributes: deposition time, depth of burial, age of the person, whether quicklime was added to the coffin at burial, whether skeleton was contaminated with oil, burial site (2 sites 130 km apart in northern England). Source: The data were collected by Jarvis (1997). The data set is electronically available from a data repository maintained by Winner, where it is listed with the name *nitrogen levels in skeletal bones of various ages and internment lengths*.

Car Objects: 93 passenger cars on sale in the United States in 1993. Criterion: sale price of the most basic version of the car. Attributes: city mileage, highway mileage, cylinders (3, 4, 5, 6, 8, rotary), engine size, maximum horsepower, engine revolutions per mile in highest gear, fuel tank capacity, passenger capacity, length, wheelbase, width, weight, rear seat room, luggage capacity, u-turn space, airbag (none, driver only, both driver and passenger), whether a manual transmission version is available, whether the manufacturer is from the United States, type of car (small, sporty, compact, midsize, large, van), drive-train type (rear, front, four-wheel drive). Source: The data set was assembled by Lock (1993) using information from *PACE New Car & Truck 1993 Buying Guide* and *Consumer Reports April 1993 Annual Auto Issue*. It is available from R package *MASS* (Venables and Ripley, 2002; Ripley et al., 2013) with label *Cars93*.

Cigarette Objects: 528 states in the USA (in different years). Criterion: packs per capita. Attributes: year, consumer price index, state population, state personal income, average state, federal, and average local excise taxes for fiscal year. Source: The data set was assembled by Professor Jonhatan Gruber, MIT. It has been used in an introductory econometrics textbook (Stock and Watson, 2003). It is available electronically from R package *Ecdat* (Croissant, 2013).

Concrete Objects: 1030 concrete samples. Criterion: concrete compressive strength. Attributes: cement (kg/m^3), blast furnace slag (kg/m^3), fly ash (kg/m^3), water (kg/m^3), superplasticizer (kg/m^3), coarse aggregate (kg/m^3), fine aggregate (kg/m^3), age in days. Source: This data set comes from a study by Yeh (1998). It is available from the UCI Machine Learning Repository (Bache and Lichman, 2013).

Contraception Objects: 210 localities in the world (most are United Nations members but includes areas like Hong Kong that are not independent countries). Criterion: percentage of unmarried women using a modern method of contraception. Attributes: annual population growth rate, per capita 2001 gross domestic product, percentage of females over the age of 15 who are economically active, population, expected number of live births per female in 2000, percentage of population that is urban in 2001. Source: The data set is reported by Weisberg (2005) who notes that the source of the data is the United Nations. It is electronically available from R package *alr3* (Weisberg, 2011) where it is labeled *UN3*.

CPU Objects: 209 central processing units on the market in 1981–1984. Criterion: published performance on a benchmark mix relative to an IBM 370/158 Model 3. Attributes: cycle time, minimum main memory, maximum main memory, cache memory, minimum number of channels, maximum number of channels. Source: The data set was assembled by Ein-Dor and Feldmesser (1987) using information from *Computerworld* magazine. It is electronically available from R package *MASS* (Venables and Ripley, 2002; Ripley et al., 2013) with label *cpus*.

Crime Objects: 47 states of the United States. Criterion: crime rate in 1960. Attributes: percentage of males aged 14–24 in state population, indicator variable for a southern state, mean years of schooling of the population aged 25 years or older, per capita expenditure on police protection in 1960, per capita expenditure on police protection in 1959, labor force participation rate of civilian urban males in the age-group 14–24, number of males per 100 females, state population in 1960, percentage of nonwhites in the population, unemployment rate of urban males 14–24, unemployment rate of urban males 35–39, wealth (median value of transferable assets or family income), income inequality (percentage of families earning below half the median income), probability of imprisonment (ratio of number of commitments to number of offenses), average time served by offenders in state prisons before their first release. Source: The data set was assembled by Ehrlich (1973) from various publications of the United States government, including *Uniform Crime Reports* of the Federal Bureau of Investigation, United States Census, and *National Prison Statistics Bulletin*. Rounded data taken from Vandaele (1978) is electronically available from OzDASL (Smyth, 2011), where it is labeled *uscrime*.

Diabetes Objects: 442 diabetes patients. Criterion: a quantitative measure of disease progression one year after baseline. Attributes: age, sex, body mass index, average blood pressure and six blood serum measurements. Source: The data was used in Efron et al. (2004). It is available electronically from R package *lars* (Hastie and Efron, 2013).

Diamond Objects: 308 round diamond stones. Criterion: sale price. Attributes: weight in carats, color purity (D, E, F, G, H, I), clarity (internally flawless, very very slight inclusion 1, very very slight inclusion 2, very slight inclusion 1, very slight inclusion 2), certification (Gemmological Institute of America, International Gemmological Institute, Hoge Raad Voor Diamant). Source: The data set was assembled by Chu (2001) from advertisements in Singapore's *Business Times* edition of February 18, 2000. It is electronically available from R package *Ecdat* (Croissant, 2013).

Dropout Objects: 63 public high schools in Chicago. Criterion: dropout rate. Attributes: enrollment, attendance rate, parental involvement rate, percent limited-English students, percent low-income students, average class size, percent White students, percent Black students, percent Hispanic students, percent Asian students, percent minority teachers, average composite ACT score, IGAP scores: reading, math, science, social science, writing. Source: This prediction problem is from a study by Czerlinski et al. (1999). Their data sources are two articles in the February 1995 issue of *Chicago* magazine (Morton, 1995; Rodkin, 1995), where the authors note that their primary data source is Illinois State Board of Education's 1994 School Report Card.

Excavator Objects: 33 hydraulic excavators operating in the opencast mining industry in the United Kingdom. Criterion: annual maintenance cost. Attributes: weight, type of machine (front shovel, backacter), type of industry (opencast coal, opencast slate), company attitude to used oil analysis (regular use, not). Source: The data are from a study by Edwards et al. (2000). The data set is electronically available from an online repository maintained by Winner, where the data set is described as *construction plant maintenance costs*.

Fish Objects: 413 female Arctic charr. Criterion: number of eggs. Attributes: age, weight, mean egg weight. Source: This prediction problem is from a study by Czerlinski et al. (1999). The data were collected by Christian Gillet from the French National Institute for Agricultural Research. The data set used in this study was obtained via personal communication in April 2012.

Fuel Objects: 51 states and the District of Columbia of the United States. Criterion: per capita motor fuel consumption in 2001. Attributes: population, fuel tax rate, per capita income, miles of federal-aid primary highways, proportion of the population who are licensed drivers. Source: The data set is reported by Weisberg (2005) who notes that the source of the data is the Federal Highway Administration. The data set is available from R package *alr3* (Weisberg, 2011) where it is labeled *Fuel2001*.

Gambling Objects: 47 British teenagers. Criterion: annual gambling expenditure. Attributes: sex, socio-economic status, weekly income, verbal score. Source: The data were collected by Ide-Smith and Lea (1988). The data set is reported by Faraway (2005) and is electronically available from associated R package *faraway* (Faraway, 2011), where it is labeled *teengamb*.

Highway Objects: 39 segments of highway in Minnesota. Criterion: accident rate. Attributes: segment length, average daily traffic count, truck volume as a percent of total volume, speed limit, number of lanes, lane width, shoulder width, number of signalized interchanges per mile, number of freeway-type interchanges per mile, number of access points per mile, highway type (federal interstate highway, principal arterial highway, major arterial, other). Source: The data set is reported by Weisberg (2005) who notes that the data were taken from an unpublished master's paper in civil engineering by Carl Hoffstedt. The data set is electronically available from R package *alr3* (Weisberg, 2011).

Hitter Objects: 322 hitters in North American Major League Baseball. Criterion: annual salary at the beginning of the 1987 season. Attributes: 1986 performance: number of at bats, hits, home runs, runs scored, runs batted in, walks, putouts, assists, errors; career performance: number of at bats, hits, home runs, runs scores, runs batted in, walks; number of years in the major leagues; division at the end of the 1986 season (East, West); league at the end of the 1986 season (American, National); league at the beginning of the 1987 season (American, National). Source: The data set was prepared by the Statistical Graphics Section of the American Statistical Association for the 1988 Annual Statistical Meetings and is available from StatLib (StatLib: Data, software and news from the statistics community). The version used in this work is from Fox (2008), includes corrections by Hoaglin and Velleman (1995), and is electronically available from a website maintained by Fox (2015).

Homeless Objects: 50 cities in the United States. Criterion: rate of homelessness. Attributes: mean temperature, unemployment rate, percentage of inhabitants with incomes below the poverty line, vacancy rate, population, percentage of public housing, whether the city has rent control. Source: The data set was assembled by Tucker (1987) from Department of Housing and Urban Development's 1984 *Report to the Secretary on the Homeless and Emergency Shelters* and other sources.

Home Objects: 3281 homes sold in San Francisco. Criterion: sales price. Attributes: number of bedrooms, interior area of the property in squarefeet, lotsize of the property, year the property was built. Source: The data were reported in Adler (2010) and are available from associated R package *nutshell* (Adler, 2012) with label *sanfrancisco.home.sales*.

Infant Objects: 105 nations. Criterion: infant-mortality rate. Attributes: percapita income, geographic location (Africa, Americas, Asia, Europe), whether the country exports oil. Source: Rates of infant mortality were obtained by Leinhardt and Wasserman (1979) from the editorial section of the *New York Times* (Crittenden, September 28 1975). The data set is reported by Fox (2008) and is electronically available from a website maintained by the author (Fox, 2015).

Laborsupply Objects: 5320 working men (in the US). Criterion: log of hourly wage. Attributes: log of annual hours worked, number of children, age, disability (yes/no), year. Source: The data comes from the Panel Study of Income Dynamics (PSID). It has been studied in Ziliak (1997). It is available electronically from R package *Ecdat* (Croissant, 2013).

Lake Objects: 69 world lakes. Criterion: number of known crustacean zooplankton species present. Attributes: surface area, maximum depth, mean depth, specific conductance, elevation, latitude, longitude, distance to nearest lake, number of lakes within 20 km, rate of photosynthesis. Source: The data set is reported by Weisberg (2005) who notes that the data were provided by Dodson and discussed in part in Dodson (1992). The data set is electronically available from R package *alr3* (Weisberg, 2011).

Land Objects: 67 counties in Minnesota. Criterion: rent per acre paid in 1977 for agricultural land planted in alfalfa. Attributes: average rent for all tillable land, density of dairy cows, proportion of pasture land, whether liming is required to grow alfalfa. Source: The data set is reported by Weisberg (2005) who notes that the data were collected by Douglas Tiffany. The data set is electronically available from R package *alr3* (Weisberg, 2011) where it is labeled *landrent*.

Lung Objects: 654 children. Criterion: forced expiratory volume in liters. Attributes: age in years, height in inches, gender, exposure to smoking. Source: The data were collected by Tager et al. (1979). The data set is reported in Ekstrom and Sørensen (2010) and is electronically available from associated R package *isdals* (Ekstrom and Sorensen, 2014) where it is labeled *fev*.

Mammal Objects: 62 mammal species. Criterion: average daily sleep. Attributes: body weight, brain weight, maximum life span, gestation time, predation index, sleep exposure index, overall danger index. Source: The data are from a study by Allison and Cicchetti (1976). The data set is available from StatLib (StatLib: Data, software and news from the statistics community), where it is labeled *sleep*.

Medical Expenditure Objects: 5574 US citizens. Criterion: annual medical expenditures. Attributes: coinsurance rate, whether the person has an individual deductible plan, log of the annual participation incentive payment, whether the person has a physical limitation, the number of chronic diseases, self-rate health (excellent, good, fair poor), log of annual family income, log of family size, years of schooling of household head, exact age, sex, child, whether household head is black. Source: The data comes from the RAND Health Insurance Experiment (RHIE). It has been studied in Deb and Trivedi (2002). It is available electronically from R package *Ecdat* (Croissant, 2013) where it is called *medexp*.

Men Objects: 34 famous men. Criterion: mean attractiveness rating. Attributes: mean likeability rating, name recognition, whether the man is American. Source: This prediction problem is from a study by Czerlinski et al. (1999). The data were collected by Henss (1996) with the participation of 115 male and 131 female Germans, in ages ranging from 17 to 66 years old.

Mileage Objects: 398 cars built in 1970–1982. Criterion: mileage. Attributes: number of cylinders, engine displacement, horsepower, vehicle weight, time to accelerate from 0 to 60 mph, model year, origin (American, European, Japanese). Source: The data set was prepared by the Committee on Statistical Graphics of the American Statistical Association for its Second Exposition of Statistical Graphics Technology, held in conjunction with the Annual Meetings in Toronto, August 15–18, 1983. It is electronically available from StatLib (StatLib: Data, software and news from the statistics community), where it is labeled *cars*. The version used in the current work is from the UCI Machine Learning Repository (Bache and Lichman, 2013), named *Auto+MPG*, in which 8 of the original cars were removed because their mileage values were missing.

Mine Objects: 44 coal mines in the Appalachian region of western Virginia. Criterion: number of fractures in upper seams of coal mines. Attributes: inner burden thickness, percent extraction of the lower previously mined seam, lower seam height, duration of operation. Source: The data set is reported by Montgomery et al. (2001) and is electronically available from associated R package *mpg* (Braun, 2012) where it is labeled *p13.7*.

Monet Objects: 430 sales of paintings by Monet. Criterion: sale price. Attributes: height of the painting, width of the painting, whether the painting is signed, auction house where sale took place. Source: The data set is reported by Greene (2003). It is electronically available from a website maintained by the author (Greene), where it is labeled *data on sales of Monet paintings*.

Mortality Objects: 60 metropolitan areas in the United States. Criterion: mortality rate. Attributes: average annual precipitation, average January temperature, average July temperature, percent population aged 65 or older, average household size, median school years completed by those over 22, percent housing units that are sound and with all facilities, humidity, population density in urbanized areas, percent nonwhite population in urbanized areas, percent employed in white collar occupations, percentage of families with income less than $3000, relative hydrocarbon pollution potential, relative nitric oxides pollution potential, relative sulfur dioxide pollution potential, annual average relative humidity. Source: The data set was assembled by McDonald and Schwing (1973). It is electronically available from StatLib (StatLib: Data, software and news from the statistics community), where it is labeled *pollution*.

Movie Objects: 62 movies. Criterion: first-run box office in the United States. Attributes: production budget, index of star poser, whether the movie is a sequel, indicator for an action film, indicator for comedy, indicator for animation, indicator for horror, MPAA rating(G, PG, PG13, R), trailer views at traileraddict.com, number of message board comments at comingsoon.net, attention at fandango.com, percentage of Fandango votes for "can't wait to see". Source: The data set is reported by Greene (2003) and is electronically available from a website maintained by the author (Greene), where it is labeled *movie buzz data*.

Mussel Objects: 44 rivers in eastern United States. Criterion: number of freshwater mussel species. Attributes: area of drainage basins, amount of dissolved solids, nitrate concentration, hydronium concentration, number of intervening rivers to four major species-source river systems: Alabama-Coosa, Apalachicola, Savannah, and St. Lawrence. Source: The data are from an article by Sepkoski and Rex (1974). The data set is electronically available from an online repository maintained by Winner, where the data set is described as *freshwater mussel species in US Rivers*.

News Objects: 39797 news articles (online). Criterion: number of shares in social networks. Attributes: 58 attributes in total, of which many are word statistics[1]. Source: The original data are from the website `www.mashable.com`. The data set was studied in

[1] For a complete description of all predictors, please see `https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity`.

Fernandes et al. (2015). It is available from the UCI Machine Learning Repository (Bache and Lichman, 2013).

Obesity Objects: 136 children. Criterion: somatotype (a scale of body type, ranging from 1, very thin, to 7, obese). Attributes: sex, body measurements at ages 2, 9, and 18: height, weight, leg circumference, strength. Source: The data were collected by Tuddenham and Snyder (1954) on children born in Berkeley, California, between January 1928 and June 1929. The data set is reported by Weisberg (2005) and is electronically available from associated R package *alr3* (Weisberg, 2011) where it is labeled *BGSall.*

Occupation Objects: 36 occupations. Criterion: prestige rating of the National Opinion Research Center (NORC). Attributes: suicide rate among males aged 20–64, median income, median number of school years completed. Source: The data set was assembled by Labovitz (1970) using data from the U.S. Census of 1950 and prestige rankings obtained by NORC in its 1947 survey. It is reported in a book by Hand et al. (1994) with identifying number 490 and label *prestige, income, education, and suicide rates for 36 occupations.*

Pinot Objects: 38 samples of Pinot Noir wine. Criterion: quality. Attributes: clarity, aroma, body, flavor, oakiness, region. Source: The data set is reported by Montgomery et al. (2001) and is electronically available from associated R package *MPV* (Braun, 2012), where it is labeled *table.b11.*

Pitcher Objects: 206 pitchers in North American Major League Baseball. Criterion: annual salary at the beginning of the 1987 season. Attributes: 1986 performance: wins, losses, earned run average, game appearances, innings pitched, games saved; career performance: wins, losses, earned run average, game appearances, innings pitched, games saved; years in major leagues; league at the end of 1986 (American, National); league at the beginning of the 1987 season (American, National). Source: The data set was prepared by the Statistical Graphics Section of the American Statistical Association for the 1988 Annual Statistical Meetings and is available from StatLib (StatLib: Data, software and news from the statistics community). The version used in this work is from Fox (2008) and is electronically available from a website maintained by the author (Fox, 2015).

Plasma Objects: 315 adults. Criterion: Plasma retinol level. Attributes: age, sex, body mass index, daily caloric intake, daily fat intake, daily fiber intake, daily cholesterol intake, dietary beta-carotene consumed per day, dietary retinol consumed per day, number of alcoholic drinks consumed per week, smoking status (never smoked, former smoker, current smoker), vitamin use (often, used but not often, not used). Source: The data set was made available by Therese Stukel, Dartmouth Hitchcock Medical Center, at StatLib (StatLib: Data, software and news from the statistics community), where it is labeled *Plasma_Retinol.* Dr. Stukel notes that a related publication is by Nierenberg et al. (1989).

Prefecture Objects: 45 prefectures in Japan. Criterion: number of emigrants to Pacific Northwest in 1911–1912 from the prefecture (per million of the prefecture's population). Attributes: percentage of land cultivated by tenant farmlands, change in ratio of tenant farmlands between 1883 and 1907, average area of arable land per farm, number of government contracted laborers sent to Hawaii, whether any of the 18 pioneer Japanese immigrants to the Pacific Northwest were from the prefecture. Source: The data are from an article by Murayama (1991). The data set is electronically available from an online repository maintained by Winner, where the data set is described as *Japanese emigration to Pacific Northwest 1880–1915.*

Prostate Objects: 97 patients with prostate cancer. Criterion: logarithm of prostate-specific antigen. Attributes: log(cancer volume), log(prostate weight), age, log(amount of benign prostatic hyperplasia), seminal vesicle invasion, log(capsular penetration), Gleason score, percentage Gleason score 4 or 5. Source: The data appears in Stamey et al. (1989). The data set is publicly available from the R package *lasso2* (Lokhorst et al., 2014), where it is labeled *Prostate.*

Reactor Objects: 32 light water reactors constructed in the United States in the late 1960s and early 1970s. Criterion: construction cost. Attributes: date on which the construction permit was issued (measured in years since January 1, 1900), time between application for and issue of the construction permit, time between issue of operating license and construction permit, net capacity, whether a prior light water reactor existed at the same site, whether the location is in the north-east region of the United States, whether a cooling tower is used, whether the nuclear steam supply system was manufactured by Babcock-Wilcox, cumulative number of power plants constructed by each architectural engineer, whether there was a partial turnkey guarantee. Source: The data set is reported by Cox and Snell (1981) and Davison (2003). It is electronically available from R package *SMPracticals* (Davison, 2013), where it is labeled *nuclear.*

Rebellion Objects: 32 Romanian counties in 1907. Criterion: proportion of villages in which rebellious events took place in the Romanian peasant rebellion of 1907, labelled *spread.* Attributes: proportion of arable land devoted to wheat, proportion of rural population that is illiterate, strength of middle peasantry (measured by the proportion of land owned in units of 7 to 50 hectares), Gini coefficient of inequality of landownership, population, region (Northern, South Central, Southwest, Eastern). Source: The data set was assembled by Chirot and Ragin (1975). Partial data set is reported by Fox (2008) and is electronically available from a website maintained by the author (Fox, 2015).

Recycle Objects: 31 Scottish local authorities. Criterion: weekly recyclate yield. Attributes: weekly recycling capacity, weekly residual capacity, number of principal materials collected, number of extended materials collected, frequency of recycling collection, frequency of residual collection, type of sort (comingled, curbside sort, dual service, single material). Source: The data were obtained by Baird et al. (2013) from Scottish local authorities. Partial data set is available electronically from an online repository maintained by Winner, where the data set is described as *recycling capacity, items collected and average yield for Scottish local authorities.*

Rent Objects: 2053 apartments in Munich, Germany. Criterion: rent per squaremeter in euros. Attributes: size, number of rooms, year of construction, whether the apartment is located at a good address, whether the apartment is located at the best address, whether the apartment has warm water, whether the apartment has central heating, whether the bathroom has tiles, whether there is special furniture in the bathroom, whether the apartment has an upmarket kitchen. Source: The data set is reported in Fahrmeir et al. (2010) and is electronically available from R package *catdata* (Schauberger and Tutz, 2014).

Salary Objects: 52 professors at a Midwestern college in the United States. Criterion: academic year salary. Attributes: sex, rank (assistant professor, associate professor, full professor), number of years in current rank, the highest degree earned (doctorate, masters), number of years since highest degree was earned. Source: The data set is reported by Weisberg (2005) and is electronically available from associated R package *alr3* (Weisberg, 2011).

SAT Objects: 50 US states. Criterion: average total score on the SAT, 1994-95. Attributes: average expenditure per pupil, average pupil to teacher ratio, average salary of teachers, percentage of eligible students. Source: The data were collected by **?**. The data set is electronically available from the R package *faraway* (Faraway, 2011).

Schooling Objects: 3010 individuals in the US. Criterion: log of wage. Attributes: lived in smsa 1966, lived in smsa in 1976, grew up near 2-yr college, grew up near 4-yr college, grew up near 4-year public college, grew up near 4-year private college, education in 1976, education in 1966, age in 1976, lived with mom and dad at age 14, single mom at 14, step parent at 14, lived in south 1966, lived in south in 1976, mom-dad education class (1-9), black, enrolled in 1976, the kww score, normed IQ score, married in 1976, library card in home at age 14, experience in 1976. Source: The data set comes from the National Longitudinal Survey of Young Men (NLSYM) and has been used by Card (1993). It is available electronically from R package *Ecdat* (Croissant, 2013).

Tip Objects: 244 parties dining in a restaurant. Criterion: tip rate. Attributes: dollar amount of the bill, size of the party, sex of the bill payer, day of the week, time of the day, whether there were smokers in the party. Source: Data were recorded by a food server in a restaurant located in a suburban shopping mall in the United States during an interval of two and a half months in early 1990. The data set is reported in a collection of case studies for business statistics (**?**). It is electronically available from R package *reshape* (Wickham, 2007).

Vote Objects: 159 counties in Georgia, USA. Criterion: proportion of uncounted votes in the 2000 presidential election. Attributes: type of voting equipment used (optical scan with central count, optical scan with precinct count, punch card, lever, paper), whether the county is in Atlanta, whether the county is urban or rural, proportion of African Americans, economic status (rich, middle, poor). Source: The data set was assembled by **?**. It is reported by Faraway (2005) and is electronically available from associated R package *faraway* (Faraway, 2011), where it is labeled *gavote*.

Wages Objects: 4360 males in the US (from 1980 to 1987). Criterion: log of wage. Attributes: year, years of schooling, years of experience, whether the wage has been set by collective bargaining, ethnicity, whether married, whether health problem, industry (12 levels), occupation (9 levels), residence (rural area, north east, nothern central, south). Source: The data set comes from the National Longitudinal Survey (NLS Youth Sample) and has been used by Vella et al. (1998). It is available electronically from R package *Ecdat* (Croissant, 2013) where it is called *Males*.

White wine Objects: 4898 white wines. Criterion: quality score (between 0 and 10). Attributes: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol. Source: This data set comes from a study by Cortez et al. (2009). It is available from the UCI Machine Learning Repository (Bache and Lichman, 2013).

References

Joseph Adler. *R in a nutshell: A desktop quick reference.* "O'Reilly Media, Inc.", 2010.

Joseph Adler. *nutshell: Data for "R in a Nutshell"*, 2012. URL `http://CRAN.R-project.org/package=nutshell`. R package version 2.0.

Truett Allison and Domenic V. Cicchetti. Sleep in mammals: Ecological and constitutional correlates. *Science*, 194(4266):732–734, 1976.

Kevin Bache and Moshe Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

Jim Baird, Robin Curry, and Tim Reid. Development and application of a multiple linear regression model to consider the impact of weekly waste container capacity on the yield from kerbside recycling programmes in scotland. *Waste Management & Research*, 31(3):306–314, 2013.

W. John Braun. *MPV: Data Sets from Montgomery, Peck and Vining's Book*, 2012. URL `http://CRAN.R-project.org/package=MPV`. R package version 1.27.

Thomas F. Brooks, D. Stuart Pope, and Michael A. Marcolini. *Airfoil self-noise and prediction*, volume 1218. National Aeronautics and Space Administration, Office of Management, Scientific and Technical Information Division, 1989. URL `https://info.aiaa.org/tac/ASG/FDTC/DG/BECAN_files_/BANCII_category1/documentation/Related_Papers/BPM-NASA-RP-1218-1989.pdf`.

David Card. Using Geographic Variation in College Proximity to Estimate the Return to Schooling. Working Paper 4483, National Bureau of Economic Research, October 1993. URL `http://www.nber.org/papers/w4483`.

Daniel Chirot and Charles Ragin. The market, tradition and peasant rebellion: The case of Romania in 1907. *American Sociological Review*, pages 428–444, 1975.

Singat Chu. Pricing the C's of diamond stones. *Journal of Statistics Education*, 9(2), 2001.

Paulo Cortez, Antnio Cerdeira, Fernando Almeida, Telmo Matos, and Jos Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009. URL `http://www.sciencedirect.com/science/article/pii/S0167923609001377`.

David R. Cox and E. Joyce Snell. *Applied Statistics: Principles and Examples.* Chapman and Hall, 1981.

Ann Crittenden. Vital dialogue is beginning between the rich and the poor. *The New York Times*, pages E–3, September 28 1975.

Yves Croissant. *Ecdat: Data sets for econometrics*, 2013. URL `http://CRAN.R-project.org/package=Ecdat`. R package version 0.2-2.

Jean Czerlinski, Gerd Gigerenzer, and Daniel G. Goldstein. How good are simple heuristics? In Gerd Gigerenzer, Peter M Todd, and the ABC Research Group, editors, *Simple heuristics that make us smart*, pages 97–118. Oxford University Press, New York, 1999.

Anthony C. Davison. *Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2003. ISBN 0-521-77339-3.

Anthony C. Davison. *SMPracticals: Practicals for use with Davison (2003) Statistical Models*, 2013. URL `http://CRAN.R-project.org/package=SMPracticals`. R package version 1.4-2.

Partha Deb and Pravin K. Trivedi. The structure of demand for health care: latent class versus two-part models. *Journal of health economics*, 21(4):601–625, 2002. URL `http://www.sciencedirect.com/science/article/pii/S0167629602000085`.

Stanley Dodson. Predicting crustacean zooplankton species richness. *Limnology and Oceanography*, 37(4):848–856, 1992.

David J. Edwards, Gary D. Holt, and Frank C. Harris. A comparative analysis between the multilayer perceptron "neural network" and multiple regression analysis for predicting construction plant maintenance costs. *Journal of Quality in Maintenance Engineering*, 6 (1):45–61, 2000.

Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, and others. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004. URL `http://projecteuclid.org/euclid.aos/1083178935`.

Isaac Ehrlich. Participation in illegitimate activities: A theoretical and empirical investigation. *Journal of Political Economy*, 81:521–565, 1973.

Phillip Ein-Dor and Jacob Feldmesser. Attributes of the performance of central processing units: A relative performance prediction model. *Communications of the ACM*, 30(4): 308–317, 1987.

Claus Ekstrom and Helle Sorensen. *isdals: Provides datasets for Introduction to Statistical Data Analysis for the Life Sciences*, 2014. URL `http://CRAN.R-project.org/package=isdals`. R package version 2.0-4.

Claus Thorn Ekstrom and Helle Sørensen. *Introduction to statistical data analysis for the life sciences*. CRC Press, 2010.

Ludwig Fahrmeir, Rita Künstler, Iris Pigeot, and Gerhard Tutz. *Statistik: Der Weg zur Datenanalyse*. Springer Berlin Heidelberg, 2010. ISBN 9783642019388.

Julian Faraway. *Extending Linear Models with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman & Hall/CRC, 2005.

Julian Faraway. *faraway: Functions and datasets for books by Julian Faraway*, 2011. URL `http://CRAN.R-project.org/package=faraway`. R package version 1.0.5.

Kelwin Fernandes, Pedro Vinagre, and Paulo Cortez. A proactive intelligent decision support system for predicting the popularity of online news. In *Portuguese Conference on Artificial Intelligence*, pages 535–546. Springer, 2015.

John Fox. *Applied regression analysis and generalized linear models.* SAGE Publications, 2nd edition, 2008.

John Fox. Applied regression analysis and generalized linear models, second edition, data sets, 2015. `http://socserv.socsci.mcmaster.ca/jfox/Books/Applied-Regression-2E/datasets/`.

William H. Greene. Econometric analysis, 7th edition, links to data tables. `http://pages.stern.nyu.edu/~wgreene/Text/econometricanalysis.htm`.

William H. Greene. *Econometric analysis.* Pearson Education India, 2003.

David J. Hand, Fergus Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski. *A handbook of small data sets.* Chapman & Hall/CRC, London, UK, 1994.

Trevor Hastie and Bradley Efron. lars: Least Angle Regression, Lasso and Forward Stagewise, April 2013.

Ronald Henss. The attractiveness of prominent people. Unpublished manuscript at the Fachrichtung Psychologie, University of Saarbrücken, Saarbrücken, Germany, 1996.

David C. Hoaglin and Paul F. Velleman. A critical look at some analyses of major league baseball salaries. *American Statistician*, 49(4266):277–285, 1995.

David Hosmer and Stanley Lemeshow. *Applied Logistic Regression.* Wiley, 2000.

Susan G. Ide-Smith and Stephen E. G. Lea. Gambling in young adolescents. *Journal of Gambling Behavior*, 4(2):110–118, 1988.

David R. Jarvis. Nitrogen levels in long bones from coffin burials interred for periods of 26–90 years. *Forensic Science International*, 85(3):199–208, 1997.

Sanford Labovitz. The assignment of numbers to rank order categories. *The American Sociological Review*, 35:515–524, 1970.

Samuel Leinhardt and Stanley S. Wasserman. Exploratory data analysis: An introduction to selected methods. *Sociological methodology*, 10:311–365, 1979.

Robin H. Lock. 1993 New car data. *Journal of Statistics Education*, 1(1), 1993.

Justin Lokhorst, Bill Venables and Berwin Turlach; port to R, and tests etc: Martin Maechler. lasso2: L1 constrained estimation aka lasso, May 2014. URL `https://cran.r-project.org/web/packages/lasso2/index.html`.

John Maindonald and W. John Braun. *Data Analysis and Graphics Using R: An Example-Based Approach.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010.

John Maindonald and W. John Braun. *DAAG: Data Analysis And Graphics data and functions*, 2013. URL `http://CRAN.R-project.org/package=DAAG`. R package version 1.16.

Gary C. McDonald and Richard C. Schwing. Instabilities of regression estimates relating air pollution to mortality. *Technometrics*, 15:463–482, 1973.

Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. Wiley, 3rd edition, 2001.

Felicia B. Morton. Charting a school's course. *Chicago*, pages 86–95, 1995.

Yuzo Murayama. Information and emigrants: Interprefectural differences of japanese emigration to the pacific northwest, 1880–1915. *The Journal of Economic History*, 51(1): 125–147, 1991.

Warwick J. Nash, Tracy L. Sellers, Simon R. Talbot, Andrew J. Cawthorn, and Wes B. Ford. The population biology of abalone (haliotis species) in Tasmania. I. Blacklip Abalone (h. rubra) from the north coast and islands of Bass Strait. *Sea Fisheries Division, Technical Report*, (48), 1994.

David W. Nierenberg, Therese A. Stukel, John A. Baron, Bradley J. Dain, and E. Robert Greenberg. Determinants of plasma levels of beta-carotene and retinol. *American Journal of Epidemiology*, 130(3):511–521, 1989.

Keith W. Penrose, A. G. Nelson, and A. G. Fisher. Generalized body composition prediction equation for men using simple measurement techniques. *Medicine & Science in Sports & Exercise*, 17(2):189, 1985.

Brian Ripley, Bill Venables, Douglas M. Bates, Kurt Hornik, Albrecht Gebhardt, and David Firth. *MASS: Support Functions and Datasets for Venables and Ripley's MASS*, 2013. URL `http://CRAN.R-project.org/package=MASS`. R package version 7.3-28.

Dennis Rodkin. 10 Keys for creating top high schools. *Chicago*, pages 78–85, 1995.

Gunther Schauberger and Gerhard Tutz. *catdata: Categorical Data*, 2014. URL `http://CRAN.R-project.org/package=catdata`. R package version 1.2.1.

J. John Sepkoski and Michael A. Rex. Distribution of freshwater mussels: coastal rivers as biogeographic islands. *Systematic Biology*, 23(2):165–188, 1974.

Jeffrey S. Simonoff. *Smoothing Methods in Statistics*. Springer Series in Statistics. Springer, 1996.

Jeffrey S. Simonoff. Online data archive. `http://people.stern.nyu.edu/jsimonof/SmoothMeth/`, 2015.

Gordon K. Smyth. Australasian Data and Story Library (OzDASL), 2011. URL `http://www.statsci.org/data`.

Robert R. Sokal and F. James Rohlf. *Biometry: The principles and practice of statistics in biological research.* W. H. Freeman and Company, San Francisco, 2nd edition, 1981.

Thomas A. Stamey, John N. Kabalin, John E. McNeal, Iain M. Johnstone, Fuad Freiha, Elise A. Redwine, and Norman Yang. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients. *The Journal of urology*, 141(5):1076–1083, 1989. URL `http://europepmc.org/abstract/med/2468795`.

StatLib: Data, software and news from the statistics community, 2013. URL `http://lib.stat.cmu.edu/`.

James H. Stock and Mark W. Watson. *Introduction to econometrics*, volume 104. Addison Wesley Boston, 2003. URL `http://www.ssc.wisc.edu/~munia/475/InstVaReg.pdf`.

Ira B. Tager, Sscott T. Weiss, Bernard Rosner, and Frank E. Speizer. Effect of parental cigarette smoking on the pulmonary function of children. *American Journal of Epidemiology*, 110(1):15–26, 1979. ISSN 0002-9262.

Richard D. Telford and Ross B. Cunningham. Sex, sport, and body-size dependency of hematology in highly trained athletes. *Medicine and Science in Sports and Exercise*, 23: 788–794, 1991.

William Tucker. Where do the homeless come from? *National Review*, pages 34–44, 1987.

Read D. Tuddenham and Margaret M. Snyder. Physical growth of California boys and girls from birth to eighteen years. *University of California Publications in child development*, 1(2):183, 1954.

Walter Vandaele. Participation in illegitimate activities: Ehrlich revisited. In A. Blumstein, J. Cohen, and D. Nagin, editors, *Deterrence and Incapacitation*, pages 270–335. National Academy of Sciences, Washington DC, 1978.

Francis Vella, Marno Verbeek, and others. Whose wages do unions raise? A dynamic model of unionism and wage rate determination for young men. *Journal of Applied Econometrics*, 13(2):163–183, 1998. URL `http://down.cenet.org.cn/upfile/54/20052918022151.pdf`.

William N. Venables and Brian D. Ripley. *Modern Applied Statistics with S.* Springer, New York, fourth edition, 2002.

Sanford Weisberg. *Applied Linear Regression.* Wiley, Hoboken NJ, 3rd edition, 2005.

Sanford Weisberg. *alr3: Data to accompany Applied Linear Regression 3rd edition*, 2011. URL `http://CRAN.R-project.org/package=alr3`. R package version 2.0.5.

Hadley Wickham. Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12), 2007. URL `http://www.jstatsoft.org/v21/i12/paper`.

Larry Winner. Miscellaneous datasets. `http://www.stat.ufl.edu/~winner/datasets.html`.

I-Cheng Yeh. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete research*, 28(12):1797–1808, 1998. URL `http://www.sciencedirect.com/science/article/pii/S0008884698001653`.

James P. Ziliak. Efficient Estimation With Panel Data When Instruments Are Predetermined: An Empirical Comparison of Moment-Condition Estimators. *Journal of Business & Economic Statistics*, 15(4):419–431, October 1997. ISSN 0735-0015, 1537-2707. doi: 10.1080/07350015.1997.10524720. URL `http://www.tandfonline.com/doi/abs/10.1080/07350015.1997.10524720`.