31

DESCRIPTION OF DATASETS

All datasets used in the examples and sample problems in this text are provided for pedagogical purposes only. They are provided so that the reader can recreate the examples included in the text. Contributors have made data available to the readers of this text on this understanding and consequently, this is the only use for which they are provided.

In some cases, datasets have been modified since the initial publication of results from the study which generated the data. In many cases, only a subset of the original data (*ie* a subset of variables or a subset of observations) are included. Consequently, the reader should not expect to be able to duplicate results obtained in the original publication.

Some of the datasets described below do not feature in examples used in the book, but may be useful as learning examples in courses using this text.

In the descriptions that follow, unless otherwise specified, all variables coded 0 or 1 (0/1) have the following meaning:

0=no, absent or negative

1=yes, present or positive

All datasets can be downloaded from the Methods in Epidemiologic Research website (http://www.upei.ca/mer).

The authors extend their sincere thanks and appreciation to the contributors of these datasets. (Individual contributors are identified and recognised in association with specific datasets.)

In addition to the datasets described in this Chapter, a number of small datasets are used for specific examples throughout the text. These are explained in sufficient detail at their time of use, so descriptions are not included in this Chapter. However, these dataset are available through the book's website.

bp **Blood pressure study**

Contributor(s)	Study type	# records	Unit of record
public domain	controlled trial	1092	bp measurement

Reference(s)

Hall S, Prescott RI, Hallman RJ, Dixon S, Harvey RE, Ball SG. A comparative study of Carvedilol, slow-release Nifedipine, and Atenolol in the management of essential hypertension. J Cardiovasc Pharmacol. 1991;18 Suppl 4:S35-8.

Brief description

A total of 1,092 diastolic blood pressure (bp) measurements were obtained from 288 patients enrolled in a clinical trial in 1 of 29 centres. Patients were randomly assigned to 1 of 3 treatments (Carvedilol, Nifedipine, Atenolol). Diastolic blood pressure was recorded at visit 1, prior to the assignment to treatment groups and the drugs' effects were evaluated at visits 3–6.

Table of variables

Variable	Description	Codes/units
centre	centre id	
patient	patient id	
visit	visit number (1–6)	
tx	treatment	1=carvedilol 2=nifedipine 3=atenolol
dbp	diastolic blood pressure	mmHg
dbp1	pre-treatment (visit #1) diastolic blood pressure	mmHg
cf	cold feet	1=none 2=occasionally 3=on most days 4=most of the time 5=all of the time
cf1	cold feet at pre-treatment visit (#1)	same coding as cf

bpwide

Brief description

A subset of the bp dataset, including only the 256 patients with complete records at all visits, is provided in wide format, whereby the values for each patient at the 5 visits are held in distinct variables as indicated in the table.

Table of variables for bpwide data

Variable	Description	Codes/units
centre, patient, tx	same as above	
dbp1, dbp3, dbp4, dbp5, dbp6	diastolic blood pressure measured at visits 1, 3, 4, 5, and 6, respectively	mmHg

brazil **Brazil water cistern study**

Contributor(s)	Study type	# records	Unit of record
Pasha Marcynuk	cross-sectional	3399	person

Reference(s)

Comparison of the burden of diarrheal illness among individuals with and without household cisterns in northeast Brazil: a cross-sectional study—submitted BMC Infectious Diseases

Cisterns, chlorine use, and other factors associated with diarrhoea among individuals living in the semi-arid region of northeast Brazil: a cross sectional study—submitted Epidemiology and Infection

Brief description

Pasha Marcynuk and colleagues at the Public Health Agency of Canada, University of Guelph, Cidade Universitária, Universidade de Pernambuco, Articulação no Semi Árido and the Pan American Health Organization kindly provided all the data from a recently completed study on the effects of the presence of a rainwater cistern at a house on the incidence of acute gastrointestinal illness (primarily diarrhea) in the semi-arid region of the Agreste Central Region of Pernambuco State in Brazil. A subset of all of the variables collected has been included (variables selected for the pedagogical value and their utility for examples in Chapters 2 and 22). The study has been classified as a cross-sectional study because data on both the outcome and the exposures of interest were collected at the same time. However, the authors retrospectively collected data over a 1-month period which allowed for the computation of incidence rates.

Table of variables

Variable	Description	Codes/units
mun	municipality	
comm	community	
fam	family	
id	individual id	
fam_n	number in family	
fam_n5	number under 5 yrs in family	
fam_nsch	number in school in family	
age	age	years
age_c8	age in 8 categories	<5, 5–10, 11–20, 21–30, 31–40, 41–50, 51–65, >65
sex	gender	0=male 1=female
diarr	diarrhea in past 30 days	
di_cnt	number of episodes of diarrhea in last month	
di_d	number of days with diarrhea in last month	
di_d_ln	days with diarrhea log transformed	

Variable	Description	Codes/units
cistern	rainwater cistern	0/1
water_tx	water treated with chlorine	0/1
water_store	water stored in clay jar	0/1
house_yrs	number of yrs owned house	years
house_rooms	number of rooms in house	
house_piped	house has piped water	0/1
house_animals~h	house has other domestic animals inside	0/1
house_garbage	method of garbage disposal	1=burned 2=left in the open 3=other
income_agric_~l	small scale agric is an important source of income	0/1
income_social	family assistance is an important source of income	0/1
income_oldage	old age pension is an important source of income	0/1
lit_mother	mother is literate	0/1
lit_father	father is literate	0/1
c_rec	number of houses in community	
c_fam_n	average family size in community	
c_age	average age in community	years
c_cistern	proportion of houses with cistern in community	
c_lit_mother	proportion of mothers literate in community	
c_lit_father	proportion of fathers literate in community	
m_pop_urban	population in urban areas in municipality	
m_pop_rural	population in rural areas in municipality	
m_pop_total	total population in municipality	
m_area_km2	area of municipality	square km
m_pop_families	number of families in municipality	
m_cistern_elig	number families eligible for cistern in municipality	
m_cistern_const	number of families with cistern constructed (2006) in municipality	
m_hdi	human development index in municipality	
m_gini	gini index in municipality	
m_cases_d_not_2006	notified cases of acute diarrhea in 2006 in municipality	
m_cases_d_not_01_06	notified cases of acute diarrhea 2001 to 2006 in municipality	
m_prpn_u14_school	prop children under 14 yrs attending school in municipality	
m_prpn_15_lit	prop people over 15 yrs literate in municipality	

Variable	Description	Codes/units
m_water_public	prop families on public water system (2006)	
m_water_well	prop families on well/spring water (2006)	
m_water_other	prop families on other water system (2006)	

brazil_smpl **Brazil water cistern study sampling weights**

Contributor(s)	Study type	# records	Unit of record
Pasha Marcynuk	cross-sectional	3399	person

Reference(s)

See preceding dataset

Brief description

A subset of the variables from the -brazil- dataset were extracted and the probabilities of a household being selected (within each of the cistern and non-cistern groups) and of an individual within the household being selected were computed in order to generate sampling weights. The study has been classified as a cross-sectional study because data on both the outcome and the exposures of interest were collected at the same time. However, the authors retrospectively collected data over a 1-month period which allowed for the computation of incidence rates.

Table of variables

Variable	Description	Codes/units
mun	municipality	
comm	community	
fam	family	
fam_n	number of people in household	
id	individual id	
cistern	water cistern	
psel_cistern	prob of household being selected within each cistern group	
psel_fam	prob of being included in the families data	
psel	overall probability of selection	
wt	sampling weight 1/psel	
age	age	years
sex	gender	0=male 1=female
diarr	diarrhea in past month	0/1
di_cnt	number of episodes of diarrhea in last month	
di_d	number of days with diarrhea in last month	
di_d_ln	days with diarrhea log transformed	
water_tx	water treated with chlorine	0/1
m_hdi	human development index	0/1
m_pop_families	number of families in municipality	

bw5k Birth-weight data

Contributor(s)	Study type	# records	Unit of record
public domain	single cohort	5000	person

Reference(s)

none

Brief description

United States' birth data for all births registered in the 50 states, the District of Columbia, and New York City were obtained from the Centers for Disease Control and Prevention's National Center for Health Statistics (NCHS) (http://www.cdc.gov/nchs/). A subset of the variables were extracted and records for which there were no missing data for those variables identified. A random sample of 5,000 observations was then selected for inclusion in this dataset. For many variables, the number of categories were reduced to avoid categories containing very few observations.

The main outcomes of interest were:
- birth weight (-bwt-) (Chapters 14 and 15)
- low birth weight (-low_bw-) defined as bwt<2500 gm (Chapters 6 and 16)
- Apgar score (apgar_c4 or apgar_c3) (Chapter 17)
- number of prenatal visits (-previs-) (Chapter 18).

Table of variables

Variable	Description	Codes/units
obs	observation number	
bwt	birth weight	gm
mrace_c4	mother's race {also as mrace_c3 (1=Hispanic, 2=White, 3=other)}	1=Hispanic 2=White 3=Black 4=other
mage	mother's age	yr
meduc_c4	mother's education	1= <high school 2= high school 3= some college 4= college degree
mar	mother's marital status	0/1
frace_c3	father's race	1=Hispanic 2=White 3=other
fage_c4	father's age	1= ≤24 2= 25–29 3= 30–34 4= ≥35
tbo	total birth order (parity)	

Variable	Description	Codes/units
previs	# of prenatal visits	
wtgain	weight gain	lb
cig_1	cigarettes smoked (per day) in 1st trimester	
cig_2	cigarettes smoked (per day) in 2nd trimester	
cig_3	cigarettes smoked (per day) in 3rd trimester	
phyper	pregnancy-associated hypertension	0/1
multbrth	multiple birth	0/1
gest	gestation length	weeks
male	sex of baby	0=female 1=male
cesarean	cesarean section	0/1
apgar	Apgar score at 5 min	1-10
apgar_c4	Apgar score recoded into 4 categories	0=1–6 1=7 2=8 3=9–10

gi_surv GI illness surveillance data

Contributor(s)	Study type	# records	Unit of record
Kate Thomas	cross-sectional	9382	person

Reference(s) (selected)

Majowicz SE, Dore K, Flint JA, Edge VL, Read S, Buffett MC, *et al*. Magnitude and distribution of acute, self-reported gastrointestinal illness in a Canadian community. Epidemiol Infect. 2004 Aug;132(4):607-17.

Majowicz SE, McNab WB, Sockett P, Henson TS, Dore K, Edge VL, *et al*. Burden and cost of gastroenteritis in a Canadian community. J Food Prot. 2006 Mar;69(3):651-9.

Majowicz SE, Horrocks J, Bocking K. Demographic determinants of acute gastrointestinal illness in Canada: a population study. BMC Public Health. 2007;7:162.

Sargeant JM, Majowicz SE, Snelgrove J. The burden of acute gastrointestinal illness in Ontario, Canada, 2005-2006. Epidemiol Infect. 2008 Apr;136(4):451-60.

Thomas MK, Majowicz SE, MacDougall L, Sockett PN, Kovacs SJ, Fyfe M, *et al*. Population distribution and burden of acute gastrointestinal illness in British Columbia, Canada. BMC Public Health. 2006;6:307.

Thomas MK, Majowicz SE, Pollari F, Sockett PN. Burden of acute gastrointestinal illness in Canada, 1999-2007: interim summary of NSAGI activities. Can Commun Dis Rep. 2008 May;34(5):8-15.

Brief description

Dr. Kate Thomas and colleagues at the Public Health Agency of Canada kindly contributed the data from surveys on the incidence of acute gastrointestinal (GI) illness carried out in three regions of Canada. A small subset of all of the variables collected (selected for their pedagogical value) have been included in this dataset. The study has been classified as a cross-sectional study because data on both the outcome and the exposures of interest were collected at the same time. However, the authors retrospectively collected data over a 1-month period which allowed for the computation of incidence rates.

Table of variables

Variable	Description	Codes/units
study	study location	1=British Columbia 2=Hamilton, Ontario 3=Ontario
region	study region	a total of 9 study regions within the 3 locations
id	participant id	
age_c20	age in 5-year increments	0=0–4 1=5–9 etc to 20=100+
sex	gender	0=female 1=male

Variable	Description	Codes/units
ill	GI disease in past month	0/1
ill_episodes	number of GI episodes in past month	
ill_days	total # of days with GI symptoms in past month	
ill_vomit	vomiting during the past month	0/1
ill_diarr	diarrhea during the past month	0/1
ill_diarr_now	currently have diarrhea	0/1

mi **Heart attack data**

Contributor(s)	Study type	# records	Unit of record
Robert Goldberg Darlene Lessard	single cohort	2965	person

Reference(s) (selected)

Floyd KC, Yarzebski J, Spencer FA, Lessard D, Dalen JE, Alpert JS, *et al*. A 30-year perspective (1975-2005) into the changing landscape of patients hospitalized with initial acute myocardial infarction: Worcester Heart Attack Study. Circ Cardiovasc Qual Outcomes. 2009 Mar;2(2):88-95.

Goldberg RJ, Glatfelter K, Burbank-Schmidt E, Lessard D, Gore JM. Trends in community mortality due to coronary heart disease. Am Heart J. 2006 Feb;151(2):501-7.

Goldberg RJ, Ciampa J, Lessard D, Meyer TE, Spencer FA. Long-term survival after heart failure: a contemporary population-based perspective. Arch Intern Med. 2007 Mar 12;167(5):490-6.

Goldberg RJ, Spencer FA, Szklo-Coxe M, Tisminetzky M, Yarzebski J, Lessard D, *et al*. Symptom presentation in patients hospitalized with acute heart failure. Clin Cardiol. 2010 Jun;33(6):E73-80.

McManus DD, Piacentine SM, Lessard D, Gore JM, Yarzebski J, Spencer FA, et al. Thirty-year (1975 to 2005) trends in the incidence rates, clinical features, treatment practices, and short-term outcomes of patients <55 years of age hospitalized with an initial acute myocardial infarction. Am J Cardiol. 2011 Aug 15;108(4):477-82.

Brief description

Dr. Goldberg and Ms. Lessard of the Department of Cardiology at the University of Massachusetts Medical School kindly extracted 2,965 records of patients from the Worcester Heart Attack Study database. These consist of patients admitted to one of 10 hospitals in the greater Worcester, Massachusetts area with an acute myocardial infarction who were followed for many years after discharge (the longest follow-up period in the data provided was slightly over 8 years). A few factors that might affect survival following the myocardial infarction were extracted and are listed below. Variables were selected for their pedagogical value and suitability for the examples in Chapter 19. A few observations with extreme values were deleted and the ages of a small number of individuals were modified to preclude any possibility of identification of individuals.

Table of variables

Variable	Description	Codes/units
id	patient id	
hosp	hospital id	
surv_mi	survival time from day of admission to hospital	days
died	observation time ended in death or censoring	0=censored 1=died
sex	gender	0=female 1=male

Variable	Description	Codes/units
age	age at admission	years
white	race	0=other 1=White
minum	mi number	
mar	marital status	1=single 2=married 3=divorced/separated 4=widowed 5=unknown 9=missing
mar_c2	married y/n	0=not married or unknown 1=married
bmi	body mass index	bmi units
prchf	previous congestive heart failure	0/1
prmi	previous myocardial infarction	0/1
prcabg	previous coronary artery bypass surgery	0/1
card	cardiac arrest during hospitalization	0/1
cath	cardiac catheterisation	0/1
ptca	coronary angioplasty	0/1
cabg	coronary artery bypass surgery	0/1
los	length of hospital stay	days
smkces	discharge education on smoking cessation	0/1

mi_mult **Heart attack data—multiple episodes**

Contributor(s)	Study type	# records	Unit of record
Robert Goldberg Darlene Lessard	single cohort	4928	myocardial infarctions

Reference(s)

See previous page

Brief description

2,230 of the patients in the -mi- dataset had complete history with regard to previous myocardial infarctions and these individuals had had a total of 4,928 episodes. This dataset contains a small number of variables used in the examples for multiple failure event data (Chapter 19).

Table of variables

Variable	Description	Codes/units
id	patient id	
age_mi	survival time from day of admission to hospital	days
mi	observation time ended in myocardial infarction or censoring	0=censored 1=mi
n	heart attack number (1st, 2nd, etc.)	
sex	gender	0=female 1=male
white	race	0=other 1=White

nv **Norovirus test data**

Contributor(s)	Study type	# records	Unit of record
David Fisman	longitudinal	188	specimen

Reference(s)

Fisman DN, Greer AL, Brouhanski G, Drews SJ. Of gastro and the gold standard: evaluation and policy implications of norovirus test performance for outbreak detection. J Transl Med. 2009;7:23.

Brief description

Dr. David Fisman kindly contributed data from a total of 188 specimens derived from outbreaks of acute gastrointestinal illness in Ontario, Canada. The specimens had been evaluated by PCR, enzyme immuno-assay (EIA) (up to 3 times) and electron microscopy (EM). Additional testing of discrepant results was also carried out (data not included in this dataset). A 'gold standard' was computed with a sample being considered positive if any of the following criteria were met:

- EM was positive
- both PCR and EIA were positive
- PCR was strongly positive (ct value <24)
- the PCR and EIA results were discrepant but either was positive when retested.

Table of variables

Variable	Description	Codes/units
sp_id	specimen id	
out_id	outbreak id	
gs	gold standard	0/1
pcr	PCR ct value	
pcr_c2	PCR neg/pos	0/1
eia1	1st EIA optical density value	
eia2	2nd EIA optical density value	
eia3	3rd EIA optical density value	
eiafin	final EIA optical density value	
eiafin_c2	EIA neg/pos	0/1
em	electron microscopy neg/pos	0/1

salm_ma ***Salmonella*** **meta-analysis data**

Contributor(s)	Study type	# records	Unit of record
Oliver Bucher	meta-analysis	381	studies

Reference(s)

Bucher O, Farrar AM, Totton SC, Wilkins W, Waddell LA, Wilhelm BJ, *et al*. A systematic review-meta-analysis of chilling interventions and a meta-regression of various processing interventions for Salmonella contamination of chicken. Prev Vet Med. 2012;103(1):1-15.

Brief description

Dr. Oliver Bucher kindly contributed the dataset used from a recently completed meta-analysis of interventions designed to reduce Salmonella levels in chicken carcasses. A small subset of the variables evaluated in the meta-analysis were selected (based on their pedagogical value). In addition, the dataset was restricted to studies carried out under commercial, or pilot scale conditions (laboratory based studies excluded). The variables selected follow.

Table of variables

Variable	Description	Codes/units
refid	reference id	
id	unique trial id	
author	senior author name	
pub_date	year of publication	
pub_type	type of publication	1=peer reviewed 4=research report
country	origin of study (grouped by continent)	5 continents
stdy_set	setting of study	1=commercial 2=pilot
stdy_dsgn	overall design of study	1=challenge trial 2=controlled trial 3=before after trial with challenge 4 = before/after trial without challenge
stdy_pop	type of birds in study	1=mixed population 2=uncontaminated birds
sple_num_cont_a	number of control birds	
sple_num_txt_a	number of treated birds	
sple_pos_cont_a	number of positives among controls	
sple_pos_txt_a	number of positives among tx	
num	total sample size	
risk_cntrl	risk in control birds	
risk_tx	risk in tx birds	
rr	risk ratio	
rr_ln	log(risk ratio)	
rrse_ln	SE of log(risk ratio)	

Variable	Description	Codes/units
conc_cont_a	Salmonella concentration in control group	
conc_txt_a	Salmonella concentration in tx group	
conc_sd_cont_a	SD of Salmonella concentration in control group	
conc_sd_txt_a	SD of Salmonella concentration in tx group	
md_ln	mean diff. in Sal. conc. (log scale)	
mdse_ln	SE of mean diff. in Sal. conc. (log scale)	
int_asn	method of intervention assignment	
blind	use of blinding techniques	0/1
intrv	type of intervention	1=scalding 2=reprocessing 3=spray 4=dip 5=chilling 6=final spray
serotype	serotype of Salm.	1=Salm. Typhimurium 2=Salmonella spp. 3=mixed spp. 4=other
disinf_c6	type of disinfectant	6 categories (see dataset)
conc_c7	concentration of chemical (7 groups)	7 categories (see dataset)
conc_c2	concentration of chemical (2 groups)	0=none 1=chemical used

salm_outbrk **Salmonella outbreak data**

Contributor(s)	Study type	# records	Unit of record
Tine Hald	matched case-control	112	individual (person)

Reference(s)

Molbak K, Hald D. An outbreak of *Salmonella* Typhimurium in the county of Funen during late summer. A case-controlled study Ugeskr Laeger. 1997; 159: 36.

Brief description

The data are from an investigation of an outbreak of Salmonella in Funen County of Denmark in 1996. The data consisted of 39 cases of *Salmonella* Typhimurium phage type 12 and 73 controls matched for age, sex and municipality of residence. Data on numerous food exposures were recorded and a small subset of those data are included in the dataset -sal_outbrk-.

Table of variables

Variable	Description	Codes/units
match-grp	case-control pair identifier	
date	interview date	
age	age	yrs
gender	gender	0 = male 1 = female
casecontrol	case-control status	0/1
eatbeef	ate beef in previous 72 hours	0/1
eatpork	ate pork in previous 72 hours	0/1
eatveal	ate veal in previous 72 hours	0/1
eatlamb	ate lamb in previous 72 hours	0/1
eatpoul	ate poultry in previous 72 hours	0/1
eatcold	ate cold sliced meats in previous 72 hours	0/1
eatveg	ate vegetables in previous 72 hours	0/1
eatfruit	ate fruit in previous 72 hours	0/1
eateggs	ate eggs in previous 72 hours	0/1
slt_a	ate pork processed at slaughterhouse A	0/1
dlr_a	ate pork marketed by wholesaler A	0/1
dlr_b	ate pork marketed by wholesaler B	0/1

vietnam (various files)

Contributor(s)	Study type	# records	Unit of record
Dirk Pfeiffer	surveillance data	134	animal groups

Reference(s)

Pfeiffer DU, Minh PQ, Martin V, Epprecht M, Otte MJ. An analysis of the spatial and temporal patterns of highly pathogenic avian influenza occurrence in Vietnam using national surveillance data Vet J. 2007; 174: 302-9.

Brief description

Data were collected through regular surveillance for outbreaks of highly pathogenic avian influenza serotype H5N1 among domestic poultry and people in a region of Northern Vietnam at the administrative level of commune between 2004 and 2006. A variety of files (some in Stata and some in R format) are provided in a zipped folder. An example of the contents of one file is shown below.

Table of variables - viet_commune_centroid

Variable	Description	Codes/units
id	commune ID	
x_coord	X coordinate of commune centroid point location	
y_coord	Y coordinate of commune centroid point location	
infected	commune infected with AI	
inf_2003_4	commune infected with AI in 2003/04	
inf_2004_5	commune infected with AI in 2004/05	
inf_2005_6	commune infected with AI in 2005/06	

Some of the other data files include:

- viet_district_centroid This geographic data layer represents commune level data aggregated at district level.
- viet_district_poly This geographic data layer represents commune level data aggregated at district level.
- viet_region_poly This geographic data layer represents the boundary of the region in Northern Vietnam for which the data was analysed.