APPENDIX A

Available data sets

A.1 Latina Mothers and their Newborn

From 1980 to 1990 data was collected on 427 Latino mothers that gave birth at the University of California, San Francisco [12, 13]. Data was collected on the characteristics of the mothers and their newborn infants (Table A.1). Mothers were weighed at each prenatal visit. Rate of weight gain during each trimester was based on a linear regression interpolation. The data set can be viewed and downloaded from `http://www.medepi.net/data/birthwt9.txt`.

Table A.1 Data dictionary for Latina mothers and their newborn infants

Variable	Description	Possible values
`age`	Maternal age	In years (self-reported)
`parity`	Parity	Count of previous live births
`gest`	Gestation	Reported in days
`sex`	Gender	Male = 1, Female = 2
`bwt`	Birth weight	Grams
`cigs`	Smoking	Number of cigarettes per day (self-reported)
`ht`	Maternal height	Measured in centimeters
`wt`	Maternal weight	Pre-pregnancy weight (self-reported)
`r1`	Rate of weight gain (1st trimester)	Kilograms per day (estimated)
`r2`	Rate of weight gain (2nd trimester)	Kilograms per day (estimated)
`r2`	Rate of weight gain (3rd trimester)	Kilograms per day (estimated)

A.2 Oswego County (outbreak)

On April 19, 1940, the local health officer in the village of Lycoming, Oswego County, New York, reported the occurrence of an outbreak of acute gastrointestinal illness to the District Health Officer in Syracuse. Dr. A. M. Rubin, epidemiologist-in-training, was assigned to conduct an investigation.

When Dr. Rubin arrived in the field, he learned from the health officer that all persons known to be ill had attended a church supper held on the previous evening, April 18. Family members who did not attend the church supper did not become ill. Accordingly, Dr. Rubin focused the investigation on the supper. He completed Interviews with 75 of the 80 persons known to have attended, collecting information about the occurrence and time of onset of symptoms, and foods consumed. Of the 75 persons interviewed, 46 persons reported gastrointestinal illness.

The onset of illness in all cases was acute, characterized chiefly by nausea, vomiting, diarrhea, and abdominal pain. None of the ill persons reported having an elevated temperature; all recovered within 24 to 30 hours. Approximately 20 physicians. No fecal specimens were obtained for bacteriologic examination.

The supper was held in the basement of the village church. Foods were contributed by numerous members of the congregation. The supper began at 6:00 p.m. and continued until 11:00 p.m. Food was spread out on table and consumed over a period of several hours. Data regarding onset of illness and food eaten or water drunk by each of the 75 persons interviewed are provided in the attached line listing (Oswego dataset). The approximate time of eating supper was collected for only about half the persons who had gastrointestinal illness.

The data set can be viewed and downloaded from `http://www.medepi.net/data/oswego.txt`. The data dictionary is provided in Table A.2 on the facing page.

A.3 Western Collaborative Group Study (cohort)

The Western Collaborative Group Study (WCGS), a prospective cohort studye, recruited middle-aged men (ages 39 to 59) who were employees of 10 California companies and collected data on 3154 individuals during the years 1960–1961. These subjects were primarily selected to study the relationship between behavior pattern and the risk of coronary hearth disease (CHD). A number of other risk factors were also measured to provide the best possible assessment of the CHD risk associated with behavior type. Additional variables collected include age, height, weight, systolic blood pressure, diastolic blood pressure, cholesterol, smoking, and corneal arcus. The median follow up time was 8.05 years.

The data set can be viewed and downloaded from `http://www.medepi.net/data/wcgs.txt`. The data dictionary is provided in Table A.3 on page 186.

Table A.2 Data dictionary for Oswego County data set

Variable	Possible values
id	Subject identificaton number
age	Age in years
sex	Sex: F = Female, M = Male
meal.time	Meal time on April 18th
ill	Developed illness: Y = Yes N = No
onset.date	Onset date: "4/18" = April 18th, "4/19" = April 19th
onset.time	Onset time: HH:MM AM/PM
baked.ham	Consumed item: Y = Yes; N = No
spinach	Consumed item: Y = Yes; N = No
mashed.potato	Consumed item: Y = Yes; N = No
cabbage.salad	Consumed item: Y = Yes; N = No
jello rolls	Consumed item: Y = Yes; N = No
brown.bread	Consumed item: Y = Yes; N = No
milk	Consumed item: Y = Yes; N = No
coffee	Consumed item: Y = Yes; N = No
water	Consumed item: Y = Yes; N = No
cakes	Consumed item: Y = Yes; N = No
vanilla.ice.cream	Consumed item: Y = Yes; N = No
chocolate.ice.cream	Consumed item: Y = Yes; N = No
fruit.salad	Consumed item: Y = Yes; N = No

A.4 Evans County (cohort)

The Evans County data set is used to demonstrate a standard logistic regression (unconditional) [15]. The data are from a cohort study in which 609 white males were followed for 7 years, with coronary heart disease as the outcome of interest.

The data set can be viewed and downloaded from `http://www.medepi.net/data/evans.txt`. The data dictionary is provided in Table A.4 on the following page.

A.5 Myocardial infarction case-control study

The myocardial infarction (MI) data set [15] is used to demonstrate conditional logistic regression. The study is a case-control study that involves 117 subjects in 39 matched strata (matched by age, race, and sex). Each stratum contains three subjects, one of whom is a case diagnosed with myocardial infarction and the other two are matched controls.

The data set can be viewed and downloaded from `http://www.medepi.net/data/mi.txt`. The data dictionary is provided in Table A.5 on page 187.

Table A.3 Data dictionary for Western Collaborative Group Study data set

Variable	Variable name	Variable type	Possible values
id	Subject ID	Integer	2001–22101
age0	Age	Continuous	39–59 years
height0	Height	Continuous	60–78 in
weight0	Weight	Continuous	78–320 lb
sbp0	Systolic blood pressure	Continuous	98–230 mm Hg
dbp0	Diastolic blood pressure	Continuous	58–150 mm Hg
chol0	Cholesterol	Continuous	103–645 mg/100 ml
behpat0	Behavior pattern	Categorical	1 = Type A1 2 = Type A2 3 = Type B1 4 = Type B2
ncigs0	Smoking	Integer	Cigarettes/day
dibpat0	Behavior pattern	Categorical	0 = Type B 1 = Type A
chd69	Coronary heart disease event	Categorical	0 = None 1 = Yes
typechd	Coronary heart disease event	Categorical	0 = CHD event 1 = Symptomatic MI 2 = Silent MI 3 = Classical angina
time169	Observation (follow up) time	Continuous	18–3430 days
arcus0	Corneal arcus	Categorical	0 = None 1 = Yes

Table A.4 Data dictionary for Evans data set

Variable	Variable name	Variable type	Possible values
id	Subject identifier	Integer	
chd	Coronary heart disease	Categorical-nominal	0 = no 1 = yes
cat	Catecholamine level	Categorical-nominal	0 = normal 1 = high
age	Age	Continuous	years
chl	Cholesterol	Continuous	> 0
smk	Smoking status	Categorical-nominal	0 = never smoked 1 = ever smoked
ecg	Electrocardiogram	Categorical-nominal	0 = no abnormality 1 = abnormality
dbp	Diastolic blood pressure	Continuous	mm Hg
sbp	Systolic blood pressure	Continuous	mm Hg
hpt	High blood pressure	Categorical-nominal	0 = no 1 = yes (dbp$\geq$ 95 or sbp$\geq$ 160)
ch	cat $\times$ hpt	Categorical	product term
cc	cat $\times$ chl	Continuous	product term

Table A.5 Data dictionary for myocardial infarction (MI) case-control data set

Variable	Variable name	Variable type	Possible values
match	Matching strata	Integer	1–39
person	Subject identifier	Integer	1–117
mi	Myocardial infarction	Categorical-nominal	0 = No 1 = Yes
smk	Smoking status	Categorical-nominal	0 = Not current smoker 1 = Current smoker
sbp	Systolic blood pressure	Categorical-ordinal	120, 140, or 160
ecg	Electrocardiogram	Categorical-nominal	0 = No abnormality 1 = abnormality

A.6 AIDS surveillance cases

`http://www.medepi.net/data/aids.txt`

A.7 Hepatitis B surveillance cases

`http://www.medepi.net/data/hepb.txt`

A.8 Measles surveillance cases

`http://www.medepi.net/data/measles.txt`

A.9 University Group Diabetes Program

`http://www.medepi.net/data/ugdp.txt`

A.10 Novel influenza A (H1N1) pandemic

A.10.1 United States reported cases and deaths as of July 23, 2009

`http://www.medepi.net/data/h1n1panflu23jul09usa.txt`