

Appendix C

The HELP study dataset

C.1 Background on the HELP study

Data from the HELP (Health Evaluation and Linkage to Primary Care) study are used to illustrate many of the entries in R and SAS. The HELP study was a clinical trial for adult inpatients recruited from a detoxification unit. Patients with no primary care physician were randomized to receive a multidisciplinary assessment and a brief motivational intervention or usual care, with the goal of linking them to primary medical care. Funding for the HELP study was provided by the National Institute on Alcohol Abuse and Alcoholism (R01-AA10870, Samet PI) and National Institute on Drug Abuse (R01-DA10019, Samet PI).

Eligible subjects were adults, who spoke Spanish or English, reported alcohol, heroin or cocaine as their first or second drug of choice, resided in proximity to the primary care clinic to which they would be referred or were homeless. Patients with established primary care relationships they planned to continue, significant dementia, specific plans to leave the Boston area that would prevent research participation, failure to provide contact information for tracking purposes, or pregnancy were excluded.

Subjects were interviewed at baseline during their detoxification stay and follow-up interviews were undertaken every 6 months for 2 years. A variety of continuous, count, discrete, and survival time predictors and outcomes were collected at each of these five occasions.

The details of the randomized trial along with the results from a series of additional analyses have been published [71, 66, 35, 47, 41, 70, 69, 79, 43, 101].

C.2 Roadmap to analyses of the HELP dataset

Table C.1 summarizes the analyses illustrated using the HELP dataset. These analyses are intended to help illustrate the methods described in the book. Interested readers are encouraged to review the published data from the HELP study for substantive analyses.

Table C.1: Analyses undertaken using the HELP dataset

Description	section (page)
Data input and output	1.13.1 (p.51)
Summarize data contents	1.13.1 (p.51)
Data display	1.13.2 (p.54)

Derived variables and data manipulation	1.13.3 (p.55)
Sorting and subsetting	1.13.4 (p.61)
Summary statistics	2.6.1 (p.78)
Exploratory data analysis	2.6.1 (p.78)
Bivariate relationship	2.6.2 (p.80)
Contingency tables	2.6.3 (p.82)
Two-sample tests	2.6.4 (p.85)
Survival analysis (logrank test)	2.6.5 (p.91)
Scatterplot with smooth fit	3.7.1 (p.111)
Linear regression with interaction	3.7.2 (p.113)
Regression diagnostics	3.7.3 (p.116)
Fitting stratified regression models	3.7.4 (p.119)
Two-way analysis of variance (ANOVA)	3.7.5 (p.120)
Multiple comparisons	3.7.6 (p.126)
Contrasts	3.7.7 (p.128)
Logistic regression	4.6.1 (p.146)
Poisson regression	4.6.2 (p.150)
Zero-inflated Poisson regression	4.6.3 (p.152)
Negative binomial regression	4.6.4 (p.154)
Quantile regression	4.6.5 (p.155)
Ordinal logit	4.6.6 (p.156)
Multinomial logit	4.6.7 (p.157)
Generalized additive model	4.6.8 (p.159)
Reshaping datasets	4.6.9 (p.160)
General linear model for correlated data	4.6.10 (p.164)
Random effects model	4.6.11 (p.166)
Generalized estimating equations model	4.6.12 (p.171)
Generalized linear mixed model	4.6.13 (p.172)
Proportional hazards regression model	4.6.14 (p.173)
Scatterplot with multiple y axes	5.6.1 (p.207)
Conditioning plot	5.6.2 (p.208)
Kaplan–Meier plot	5.6.3 (p.209)
ROC curve	5.6.4 (p.211)
Pairs plot	5.6.5 (p.213)
Visualize correlation matrix	5.6.6 (p.214)
Multiple imputation	6.5 (p.228)
Bayesian Poisson regression	6.6 (p.231)
Cronbach α	6.7.1 (p.233)
Factor analysis	6.7.2 (p.234)
Recursive partitioning	6.7.3 (p.237)
Linear discriminant analysis	6.7.4 (p.238)
Hierarchical clustering	6.7.5 (p.240)

C.3 Detailed description of the dataset

The Institutional Review Board of Boston University Medical Center approved all aspects of the study, including the creation of the de-identified dataset. Additional privacy protection was secured by the issuance of a Certificate of Confidentiality by the Department of Health and Human Services.

A de-identified dataset containing the variables utilized in the end of chapter examples is available for download at the book website:

<http://www.math.smith.edu/sasr/datasets/help.csv>.

Variables included in the HELP dataset are described in Table C.2. A copy of the study instruments can be found at: <http://www.math.smith.edu/help>.

Table C.2: Annotated description of variables in the HELP dataset

VARIABLE	DESCRIPTION	VALUES	NOTE
a15a	number of nights in overnight shelter in past 6 months	0–180	see also homeless
a15b	number of nights on the street in past 6 months	0–180	see also homeless
age	age at baseline (in years)	19–60	
anysubstatus	use of any substance post-detox	0=no, 1=yes	see also daysanysub
cesd*	Center for Epidemiologic Studies Depression scale	0–60	see also f1a–f1t
d1	how many times hospitalized for medical problems (lifetime)	0–100	
daysanysub	time (in days) to first use of any substance post-detox	0–268	see also anysubstatus
daysdrink	time (in days) to first alcoholic drink post-detox	0–270	see also drinkstatus
dayslink	time (in days) to linkage to primary care	0–456	see also linkstatus
drinkstatus	use of alcohol post-detox	0=no, 1=yes	see also daysdrink
drugrisk*	Risk-Assessment Battery (RAB) drug risk score	0–21	see also sexrisk
e2b*	number of times in past 6 months entered a detox program	1–21	
f1a	I was bothered by things that usually don't bother me	0–3 [#]	
f1b	I did not feel like eating; my appetite was poor	0–3 [#]	
f1c	I felt that I could not shake off the blues even with help from my family or friends	0–3 [#]	
f1d	I felt that I was just as good as other people	0–3 [#]	
f1e	I had trouble keeping my mind on what I was doing	0–3 [#]	
f1f	I felt depressed	0–3 [#]	
f1g	I felt that everything I did was an effort	0–3 [#]	
f1h	I felt hopeful about the future	0–3 [#]	
f1i	I thought my life had been a failure	0–3 [#]	

f1j	I felt fearful	0-3 [#]	
f1k	My sleep was restless	0-3 [#]	
f1l	I was happy	0-3 [#]	
f1m	I talked less than usual	0-3 [#]	
f1n	I felt lonely	0-3 [#]	
f1o	People were unfriendly	0-3 [#]	
f1p	I enjoyed life	0-3 [#]	
f1q	I had crying spells	0-3 [#]	
f1r	I felt sad	0-3 [#]	
f1s	I felt that people dislike me	0-3 [#]	
f1t	I could not get going	0-3 [#]	
female	gender of respondent	0=male, 1=female	
g1b*	experienced serious thoughts of suicide (last 30 days)	0=no, 1=yes	
homeless*	1 or more nights on the street or shelter in past 6 months	0=no, 1=yes	see also a15a and a15b
i1*	average number of drinks (standard units) consumed per day (in the past 30 days)	0-142	see also i2
i2	maximum number of drinks (standard units) consumed per day (in the past 30 days)	0-184	see also i1
id	random subject identifier	1-470	
indtot*	Index of Drug Abuse Consequences (InDuc) total score	4-45	
linkstatus	post-detox linkage to primary care	0=no, 1=yes	see also dayslink
mcs*	SF-36 Mental Composite Score	7-62	see also pcs
prec*	number of primary care visits in past 6 months	0-2	see also linkstatus , not observed at baseline
pcs*	SF-36 Mental Composite Score	14-75	see also mcs
pss_fr	perceived social supports (friends)	0-14	see also dayslink
satreat	any BSAS substance abuse treatment at baseline	0=no, 1=yes	
sexrisk*	Risk-Assessment Battery (RAB) drug risk score	0-21	see also drugrisk
substance	primary substance of abuse	alcohol, cocaine or heroin	
treat	randomization group	0=usual care, 1=HELP clinic	

Notes: Observed range is provided (at baseline) for continuous variables.

* denotes variables measured at baseline and followup (e.g., **cesd** is baseline measure, **cesd1** is measure at 6 months, and **cesd4** is measure at 24 months).

[#]: For each of the 20 items in HELP section F1 (CESD), respondents were asked to indicate

how often they behaved this way during the past week (0 = rarely or none of the time, less than 1 day; 1 = some or a little of the time, 1-2 days; 2 = occasionally or a moderate amount of time, 3-4 days; or 3 = most or all of the time, 5-7 days); items **f1d**, **f1h**, **f1l** and **f1p** were reverse coded.