**Stata Data-File Documentation for Smithson, M. & Merkle, E.C. (2013).**
*Generalized Linear Models for Categorical and Continuous Limited Dependent Variables.*
**Chapman & Hall.**

**ch2_skip.dta**
Description:
Data from the U.S. National Survey on Drug Use and Health on the frequency with which individuals skip school and other covariates.

Format:
A .dta file with 252 observations on the following 6 variables.

| | |
|---|---|
| income | Reported household income, where 1 means < $20k; 2 means >= $20k and < $50k; 3 means >= $50k and < $75k; 4 means >= $75k. |
| irsex | Gender; 1 is male and 2 is female. |
| educatn2 | Grade in school (see details). |
| schdskip | Reported number of school days skipped out of the past 30. |
| wrkhrsw2 | Reported number of hours worked in the past week. |
| anyskip | A binary version of schdskip, signifying whether the respondent skipped any days of school out of the past 30. |

Details:
Variable names match those from the National Survey on Drug Use and Health, so more details can be obtained from the survey codebook. Missing data codes have been changed to NA. Additionally, the educatn2 has been recoded to generally match the actual grade in which the respondent is enrolled. The only exceptions to this are that 14 means the second and third years in college, and 15 means the fourth or higher year in college.

Source:
Obtained from the Inter-University Consortium for Political and Social Research, University of Michigan, http://www.icpsr.umich.edu.

References:
United States Department of Health and Human Services. Substance Abuse and Mental Health Services Administration. Center for Behavioral Health Statistics and Quality. National Survey on Drug Use and Health, 2010. ICPSR32722-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2011-12-05. doi:10.3886/ICPSR32722.v1

**ch3_babies.dta**
Description:
Gaze patterns of four babies in a group.

Format:
A .dta file with 1180 observations on the following variables.

| | |
|---|---|
| row | a numeric vector |
| time | a numeric vector indexing the target baby |
| id | a numeric vector indexing the observations |
| gaze | a factor indicating whether a baby was looked at, with levels no yes |
| babies | a factor indexing which baby was chosen to be looked at with levels baby1 baby2 baby3 baby4 |
| lookedat | a numeric vector registering whether gaze was initiated by the target baby, with levels 0 indicating "no" and 1 indicating "yes" |
| numgaze | equals 1 if the baby is the one chosen to be looked at and 0 otherwise |
| babe1 | equals 1 if it is baby 1 and 0 otherwise |
| babe2 | equals 1 if it is baby 2 and 0 otherwise |
| babe3 | equals 1 if it is baby 3 and 0 otherwise |
| babe4 | equals 1 if it is baby 4 and 0 otherwise |
| lookbabe1 | product of lookedat and babe1 |
| lookbabe2 | product of lookedat and babe2 |
| lookbabe3 | product of lookedat and babe3 |

lookbabe4   product of lookedat and babe4

Source:
These are hypothetical data.

## ch3_cocaine.dta
Description:
Data from the 1991-1994 Drug Abuse Treatment Outcome Study on cocaine usage patterns.

Format:
A .dta file with 7592 observations on the following 2 variables.
sex         a factor with levels female male
mode        a factor recording self-reported method of cocaine ingestion with levels crack freebase
            inhale inject

Source:
The data are extracted from the 1991-1994 Drug Abuse Treatment Outcome Study (2010) data
(DATOS).

References:
United States Department of Health and Human Services and National Institute of Health and
National Institute on Drug Abuse (2010). Drug Abuse Treatment Outcome Study, 1991-1994. Ann
Arbor, MI: Inter-university Consortium for Political and Social Research.

## ch3_cocaineplus.dta
Description:
Data from the 1991-1994 Drug Abuse Treatment Outcome Study on cocaine usage patterns.

Format:
A data frame with 7592 observations on the following 8 variables.
sexsrt      a factor with levels FEMALE MALE
age         a numeric vector
mstatstr    a factor with levels BLANK DIVORCED LIVINGASMARRIED MARRIED
            NEVERMARRIED SEPARATED WIDOWED
modestr     a factor with levels crack freebase inhale inject
racestr     a factor with levels AfroAmerican Caucasian Hispanic Other
sex         a numeric vector that takes the value 1 if male and 0 if female
mode        a numeric vector that takes the value 1 if cocaine usage method is crack, 2 if method is
            freebase, 3 if method is inhale, and 4 if method is inject
race        a numeric vector that takes the value 1 if AfroAmerican, 2 if Caucasian, 3 if Hispanic, and
            4 if Other

Source:
The data are extracted from the 1991-1994 Drug Abuse Treatment Outcome Study (2010) data
(DATOS).

References:
United States Department of Health and Human Services and National Institute of Health and
National Institute on Drug Abuse (2010). Drug Abuse Treatment Outcome Study, 1991-1994. Ann
Arbor, MI: Inter-university Consortium for Political and Social Research.

**ch3_trchoice2.dta**

Description:
Choice of transportation mode by gender, in long format so that each choice occupies 5 rows.

Format:
A .dta file with 50 rows on the following variables.

| | |
|---|---|
| case | a numeric vector |
| sex | a numeric vector, = 1 if male and 0 if female |
| resp | a factor indicating the transport mode choice, where B denotes taking the bus, D denotes driving one's own vehicle, F denotes getting a ride with a friend, O denotes other, and W denotes walking |
| chosen | a numeric vector taking the value 1 if the transport mode was chosen and 0 if not |
| fr | a numeric vector recording the frequency of the case |
| pubpriv | a numeric vector that takes a value of 1 if the transportation mode is private and 0 if it is public |
| numresp | a numeric vector corresponding to transport choice |
| walk | a binary variable that equals 1 if walking was chosen and 0 otherwise |
| drive | a binary variable that equals 1 if driving was chosen and 0 otherwise |
| friend | a binary variable that equals 1 if friend driving was chosen and 0 otherwise |
| bus | a binary variable that equals 1 if bus was chosen and 0 otherwise |
| other | a binary variable that equals 1 if other was chosen and 0 otherwise |

Source:
The data are extracted from the 1991-1994 Drug Abuse Treatment Outcome Study (2010) data (DATOS).

References:
United States Department of Health and Human Services and National Institute of Health and National Institute on Drug Abuse (2010). Drug Abuse Treatment Outcome Study, 1991-1994. Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

**ch4_email.dta**

Description:
Data from the U.S. General Social Surveys on marital status (ordinal; see details) and email usage.

Format:
A .dta file with 3967 observations on the following 3 variables.

| | |
|---|---|
| marital | Marital status, an ordered factor with levels never.married < married < divorced. |
| email.hrs | Reported weekly hours spent emailing. |
| z.email | Standardized version of email.hrs. |

Details:
In creation of this dataset, an additional GSS item (DIVORCE) was used to ensure that married people in the sample had not been previously divorced or widowed. Thus, the marital status variable in this dataset is truly ordinal, as individuals can only progress through the statuses in one order.

Source:
The Survey Documentation and Analysis system hosted at UC, Berkeley:
http://sda.berkeley.edu/GSS/.

References:
Smith, T. W., Marsden, P. V., Hout, M., & Kim, J. (2011). General Social Surveys, 1972 - 2010. Principal Investigator, Tom W. Smith; Co-Principal-Investigators, Peter V. Marsden and Michael Hout, NORC ed. Chicago: National Opinion Research Center, producer, 2005; Storrs, CT: The Roper Center for Public Opinion Research, University of Connecticut, distributor. 1 data file (55,087 logical records) and 1 codebook (3,610 pp).

**ch4_treatment.dta**
Description:
Experimental data in which participants were presented with statistical information about two treatments for chest pain, then asked about their preference for a treatment.

Format:
A .dta file with 235 observations on the following 4 variables.
cond        Condition, referring to the way that statistical information was presented (see details).
choice      Treatment preference on an ordinal, 6-level scale from "definitely angioplasty" to "definitely bypass".
effectiveness    Participant ratings of the importance of treatment effectiveness on treatment choice (1 is extremely unimportant; 6 is extremely important).
invasiveness    Participant ratings of the importance of treatment invasiveness on treatment choice (1 is extremely unimportant; 6 is extremely important).

Details:
The data were taken from Hulsey (2010). Study participants were asked to make a hypothetical decision between two treatments for chest pain: bypass surgery or balloon angioplasty. Bypass is generally more effective, but it is also more invasive and has a longer recovery time.
Conditions were defined by the way participants received statistical information concerning the two treatments. In condition pictograph, participants viewed visual information via a pictograph. In condition statistics, participants view numerical information.

Source:
Provided by Lukas Hulsey.

References:
Hulsey, L. (2010). Testimonials and statistics in patient medical decision aids. Unpublished master's thesis, Wichita State University.

**ch5_fixation.dta**
Description:
Summary eyetracking data from a study examining the impact of text saliency on eye movements.

Format:
A .dta file with 48 observations on the following 6 variables.
id          Participant ID label.
condition   Condition, signifying whether a channel had a red title (see details).
countleft   Count of fixations in the middle, left channel.
countright  Count of fixations in the middle, right channel.
gazetime    Total gaze time on the webpage.
rt.cond     Equals red if the middle, right channel title was red; black otherwise.

Details:
The data are taken from Owens, Shrestha, & Chaparro (2009). A webpage was divided into 9 channels (sections), and the title color of the "middle, left" and "middle, right" channels were manipulated.
The variable condition takes the value Control if all title colors were black; Left if the "middle, left" channel title was red; and Right if the "middle, right" channel title was red.

Source:
Provided by Justin W. Owens.

References:
Owens, J.W., Shrestha, S., & Chaparro, B.S. (2009). Effects of text saliency on eye movements while browsing a web portal. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 53, pp. 1257-1261).

**ch5_workdays.dta**

Description:

Data from the U.S. National Survey on Drug Use and Health on the frequency with which individuals miss work due to mental health issues and other covariates.

Format:

A .dta file with 777 observations on the following 8 variables.

| | |
|---|---|
| cigtry | Reported age that the respondent first smoked a cigarette. |
| impydays | Reported days in the past year the respondent was unable to work due to mental health (see details). |
| age2 | Respondent age (see details). |
| service | Has the respondent been in the U.S. Armed Forces? (1=yes, 2=no) |
| health | Rating of overall health, where 1 is excellent and 5 is poor. |
| movespy2 | Number of times the respondent moved in the past 12 months. |
| schenrl | Whether the respondent is enrolled in any school (1=yes, 0=no). |
| coutyp2 | Type of county in which the respondent resides: large metro (large), small metro (small), nonmetro (nonmetro). |

Details:

Variable names match those from the National Survey on Drug Use and Health, so more details can be obtained from the survey codebook. Missing data codes have been changed to NA. Additionally, age2 is coded so that 7 means 18 years of age, 8 means 19 years of age, ..., 11 means 22 or 23 years of age, 12 means 24 or 25 years, 13 means 26-29, 14 means 30-34, 15 means 35-49, 16 means 50-64, and 17 means 65 and over.

The variable impydays contains responses to the question "About how many days out of 365 in the past 12 months were you totally unable to work or carry out your normal activities because of your emotions, nerves, or mental health?"

Source:

Obtained from the Inter-University Consortium for Political and Social Research, University of Michigan, http://www.icpsr.umich.edu.

References:

United States Department of Health and Human Services. Substance Abuse and Mental Health Services Administration. Center for Behavioral Health Statistics and Quality. National Survey on Drug Use and Health, 2010. ICPSR32722-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2011-12-05. doi:10.3886/ICPSR32722.v1

**ch6_carsales.dta**

Description:

Replication of the car salesperson problem in See, Fox, and Rottenstreich (2006).

Format:

A .dta file with 155 observations on the following 4 variables.

| | |
|---|---|
| initial | a numeric vector taking the value 0 for the Car condition and 1 for the Salesperson condition |
| prob | a numeric vector recording the respondent's probability estimate that the car was purchased from Carlos |
| NFCC | a numeric vector recording respondents' scores on the Need for Certainty and Closure scale |
| ctrNFCC | a numeric vector that is NFCC standardized to have a mean of 0 and standard deviation of 1 |
| initnfcc | a numeric vector consisting of the product of initial and ctrNFCC |

Source:

Provided by Gurr, M. (2009).

References:

Gurr, M. (2009). Partition dependence: Investigating the principle of insufficient reason, uncertainty and dispositional predictors. (Unpublished Honours thesis: The Australian National University,

Canberra, Australia)
See, K. E., Fox, C. R., & Rottenstreich, Y. S. (2006). Between ignorance and truth: Partition dependence and learning in judgment under uncertainty. Journal of Experimental Psychology, 32, 1385-1402.

## ch6_dyslexic3.dta

Description:
Reading scores and nonverbal IQ scores for gender- and age-matched dyslexic and non-dyslexic readers.

Format:
A .dta file with 44 observations on the following 4 variables.

score      a numeric vector recording childrens' scores on a reading accuracy test
dys        a numeric vector taking the value 1 if dyslexic and 0 if not
ziq        a numeric vector recording childrens' nonverbal IQ scores, standardized to have a mean of 0 and standard deviation of 1
tscore     score with all 1's replaced by.99

Details:
The reading accuracy scores have a maximum score of 1, indicating a perfect score on the test. In the Example 6.2 analysis, these are recoded to .99 (via the tscore variable); whereas in the 1's inflated model in Ch. 6 and the censored regression model in Ch. 7 they have a value of 1.

Source:
Provided by Pammer and Kevan (2007), first analyzed in Smithson and Verkuilen (2006).

References:
Pammer, K., & Kevan, A. (2007). The contribution of visual sensitivity, phonological processing, and nonverbal IQ to childrens reading. Scientific Studies in Reading, 11, 33-53.
Smithson, M. J., & Verkuilen, J. (2006). A better lemon squeezer? maximum likelihood regression with beta-distributed dependent variables. Psychological Methods, 11, 54-71.

## ch6_guilt1.dta

Description:
Judged probabilities of guilt in a criminal trial scenario (Study 1).

Format:
A .dta file with 104 observations on the following 7 variables.

observ     a numeric vector indexing cases
crguilt    a numeric vector recording the judged probability of guilt in a criminal trial scenario
cigult     a numeric vector recording the judged probability of guilt in a civil trial scenario
crvd1      a numeric vector taking the value 1 if the respondent returned a "guilty" verdict in the criminal trial and 0 otherwise
crvd2      a numeric vector taking the value 1 if the respondent returned a "not guilty" verdict in the criminal trial and 0 otherwise
civd1      a numeric vector taking the value 1 if the respondent returned a "guilty" verdict in the civil trial and 0 otherwise
civd2      a numeric vectortaking the value 1 if the respondent returned a "not guilty" verdict in the civil trial and 0 otherwise

Source:
Data provided from Study 1 of Smithson, Deady and Gracik (2007).

References:
Smithson,M., Gracik, L., & Deady, S. (2007). Guilty, not guilty, or …? multiple verdict options in jury verdict choices. Journal of Behavioral Decision Making, 20, 481-498.

**ch6_guilt3.dta**
Description:
Judged probabilities of guilt in a criminal trial scenario (Study 3).

Format:
A .dta file with 96 observations on the following 3 variables.
pguilt    a numeric vector recording the judged probability of guilt in a criminal trial scenario
v1        a numeric vector taking the value 1 if the respondent returned a "guilty" verdict in the
          criminal trial and 0 otherwise
v2        a numeric vector taking the value 1 if the respondent returned a 'not 'guilty" verdict in the
          criminal trial and 0 otherwise

Source:
Data provided from Study 3 of Smithson, Deady and Gracik (2007).

References:
Smithson,M., Gracik, L., & Deady, S. (2007). Guilty, not guilty, or …? multiple verdict options in jury
verdict choices. Journal of Behavioral Decision Making, 20, 481-498.

**ch7_dass.dta**
Description:
Depression, Anxieity, and Stress Scale Data.

Format:
A .dta file with 166 observations on the following 3 variables.
depress   a numeric vector measuring depression, scored from 0 to 20
anxiety   a numeric vector measuring anxiety, scored from 0 to 20
stress    a numeric vector measuring stress, scored from 0 to 20

Source:
Data from a pilot study by Michael Smithson.

References:
Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison
of the Depression Anxiety Stress Scales with the Beck Depression and Anxiety Inventories.
Behavior Research and Therapy, 33, 335-343.

**ch7_euthan.dta**
Description:
Euthanasia scale and Christian identification scale data.

Format:
A .dta file with 351 observations on the following 3 variables.
mident    a numeric vector measuring the degree to which respondents identify themselves as
          Christian, on a scale from 0 to 1
teuth     a numeric vector measuring the degree to which respondents favor euthanasia, on a
          scale from 0 to 1
status    a numeric vector taking the value 0 if the observation is censored and 1 if not

Source:
Data obtained from Mavor's (2004) study.

References:
Mavor, K. (2004). Religious orientation, social identity and attitudes to homosexuality. Unpublished
doctoral dissertation, School of Psychology, The Australian National University, Canberra,
A.C.T., Australia.

**ch7_exam.dta**

Description:

Grades achieved by second-year psychology students at The Australian National University in an introductory research methods course and the percentage marks they received in the laboratory component of that course.

Format:

A .dta file with 154 observations on the following 3 variables.

| | |
|---|---|
| labs | a numeric vector recording the percentage mark for the laboratory component of the course |
| final | a numeric vector recording the percentage mark for the final exam |
| cens | a numeric vector taking the value 100 to indicate the value of censored observations |

Source:

Data obtained from Michael Smithson.

**ch7_grades.dta**

Description:

Lab percentage mark, letter grade, lower and upper grade thresholds, a censored variable value, and the final percentage course mark.

Format:

A .dta file with 165 observations on the following 6 variables.

| | |
|---|---|
| lab | a numeric vector recording the percentage mark for the laboratory component of the course |
| gradecat | a factor denoting the letter grade for the course, with levels CR D HD N P |
| lower | a numeric vector denoting the lower threshold for the corresponding letter grade |
| upper | a numeric vector denoting the upper threshold for the corresponding letter grade |
| cens | a numeric vector listing the censoring value of a mark, 3 |
| finmark | a numeric vector recording the final percentage mark for the course |

Source:

Data obtained from Michael Smithson.

**ch7_rtime.dta**

Description:

Response times for a task timed-out at 1200 ms, and a prime (either respondents were primed to use intuition or deliberation in the task).

Format:

A .dta file with 300 observations on the following 3 variables.

| | |
|---|---|
| RT | a numeric vector, response time in milliseconds |
| prime | a numeric vector taking the value 0 if primed to use intuition or 1 if primed to use deliberation |
| status | a numeric vector taking the value 0 if the observation is censored and 1 if not |

Source:

These are hypothetical data.

**ch7_intervalbeta.dta**

Description:

Lower and upper probability estimates provided by the Busdecu et al. (2009) respondents in their interpretations of the phrase "very likely" in an IPCC report statement, along with dummy variables indicating the experimental condition.

Format:

A .dta file with 220 observations on the following 5 variables.

t       a numeric vector taking the value 1 if the respondent is in the Translation condition, and 0
        otherwise
n       a numeric vector taking the value 1 if the respondent is in the Narrow condition, and 0
        otherwise
w       a numeric vector taking the value 1 if the respondent is in the Wide condition, and 0 otherwise
y1      a numeric vector recording the respondent's lower probability estimate
y2      a numeric vector recording the respondent's upper probability estimate

Source:
Data provided by D. V. Budescu from the Budescu et al. (2009) study.

References:
Budescu, D.V., Broomell, S., and Por,H.-H. (2009). Improving the communication of uncertainty in
the reports of the Intergovernmental panel on climate change, Psychological Science, 20, 299-308.