ARE213 Problem Set #1A

Frank Proulx and Peter Alstone

September 19, 2013

1 Problem #1

- Data records are excluded from the dataset based whether the following variables take the noted values *as found in the data manual*:
- One easy way that the data cleaning process could be improved would be to only remove records based on the variables of interest. Missing values in fields that are not actually being used in the analysis do not pose a problem.
- Summary table:

2 Problem #2

a The mean differences in

b The average treatment effect of maternal smoking can be determined by comparing the unadjusted difference in mean birth weight of infants if we have randomized control and treatment groups of mothers.

summarytable	var	n	mean	sd	se
rectype	1	114616	1.26	0.44	0.00
pldel3	2	114616	1.02	0.13	0.00
birattnd	3	114616	1.20	0.56	0.00
cntocpop	4	114616	1.44	1.14	0.00
stresfip	5	114616	41.74	2.17	0.01
dmage	6	114616	27.76	5.70	0.02
ormoth	7	114616	0.09	0.52	0.00
mrace3	8	114616	1.26	0.66	0.00
dmeduc	9	114616	13.21	2.27	0.01
dmar	10	114616	1.25	0.43	0.00
adequacy	11	114616	1.30	0.55	0.00
nlbnl	12	114616	0.97	1.15	0.00
dlivord	13	114616	1.99	1.17	0.00
dtotord	14	114616	2.42	1.52	0.00
totord9	15	114616	2.41	1.46	0.00
monpre	16	114616	2.50	1.33	0.00
nprevist	17	114616	11.15	3.52	0.01
disllb	18	114616	350.42	362.33	1.07
isllb10	19	114616	3.32	3.19	0.01
dfage	20	114616	30.06	6.41	0.02
orfath	21	114616	0.09	0.53	0.00
dfeduc	22	114616	13.28	2.33	0.01
birmon	23	114616	6.47	3.39	0.01
weekday	24	114616	4.05	1.88	0.01
dgestat	25	114616	39.15	2.44	0.01
csex	26	114616	1.49	0.50	0.00
dbrwt	27	114616	3373.30	585.17	1.73
dplural	28	114616	1.03	0.17	0.00
omaps	29	114616	8.12	1.26	0.00
fmaps	30	114616	9.01	0.71	0.00
clingest	31	114616	39.11	2.06	0.01
delmeth5	32	114616	1.55	1.01	0.00
anemia	33	114616	1.99	0.10	0.00
cardiac	34	114616	1.99	0.08	0.00
lung	35	114616	1.99	0.08	0.00
diabetes	36	114616	1.97	0.16	0.00
herpes	37	114616	1.99	0.09	0.00
chyper	38	114616	1.99	0.09	0.00
phyper	39	114616	1.97	0.17	0.00
pre4000	40	114616	1.99	0.12	0.00
preterm	41	114616	1.99	0.12	0.00
tobacco	42	114616	1.84	0.37	0.00
cigar	43	114616	1.91	5.30	0.02
cigar6	44	114616	0.35	0.86	0.00
alcohol	45	114616	1.99	0.10	0.00

smoketable	Mean Value (Smokers)	Mean Value (Non-Smokers)	Mean Differe
1 minute APGAR score	8.10275922478923	8.12019719771666	1.7437972927432
5 munute APGAR score	9.00908792291689	9.00922677737416	1.3885445726202
birthweight	3171.13916566298030	3411.61984431759220	2.4048067865461

3 Appendix

R code for problem #1:

```
### This is Frank Proulx’s solution to ARE213 PS1a, problem 1
## Data is in the file "ps1.dta"

library(foreign) #this is to read in Stata data
library(Hmisc)
library(psych)
data <- read.dta("ps1.dta")

print(nrow(data))

## Problem 1a: Fix missing values
## The following are the error codes for each of the 15 variables that need fixing:
# cardiac: 9
# lung: 9
# diabetes: 9
# herpes: 9
# chyper: 9
# phyper: 9
# pre4000: 9
# preterm: 9
# tobacco: 9
# cigar: 99
# cigar6: 6
# alcohol: 9
# drink: 99
# drink5: 5
# wgain: 99

data <- subset (data, (cardiac != 9) & (lung != 9) & (diabetes !=9) & (herpes !=9)

print(nrow(data)) #number of records remaining after cleaning
print(describe(data, skew=FALSE, ranges=FALSE))

write.csv(data, file = "ps1dataclean.csv")

#’omaps’ and ’fmaps’ are the APGAR scores
#’dbrwt’ is the birth weight in grams
# ’tobacco’ is smoker status (1=yes, 2=no)

smokers <- subset(data, tobacco==1)
nonsmokers <- subset(data, tobacco==2)

smokerstats <- c(mean(smokers$omaps), mean(smokers$fmaps), mean(smokers$dbrwt))
nonsmokerstats <- c(mean(nonsmokers$omaps), mean(nonsmokers$fmaps), mean(nonsmokers
meandif <- nonsmokerstats - smokerstats

print(smokerstats)
print(nonsmokerstats)
print(meandif)

print(t.test(data$omaps~data$tobacco))
print(t.test(data$fmaps~data$tobacco))
print(t.test(data$dbrwt~data$tobacco))
```