ARE213 Problem Set #1A

Peter Alstone & Frank Proulx

September 25, 2013

1 Problem #1

1.1 Part A

Data records are excluded from the dataset based whether the following variables take the noted values as found in the data manual (see code for implementation).

1.2 Part B

We dropped all rows where any data were missing in that row. One way that the data cleaning process could be improved would be to only remove records based on the variables of interest (as are determined in subsequent analysis) since missing values in fields that are not eventually used in the analysis do not pose a problem.. This would result in a more iterative approach, however, and increase workload on the researcher.

We used some exploratory analysis to understand if the records that were dropped due to missing data *somewhere* in the record were representative. First we compared a few simple summary statistics between the "full record" and "partial record" data on variables of interest for this analysis. These are summarized in Table 1. Better APGAR scores and lower incidence of smoking may be correlated with having full datasets, which indicates the people who have missing data may bias the sample. We also used agnostic linear regression to understand the relationship between the presence of full records and three key variables: one-minute apgar (omaps), five-minute apgar (fmaps), and number of cigarettes smoked each day (cigar). The results summarized in Table 2 indicate there is statistical significance in each of the

factors (i.e. all three are useful predictors for whether a person has a full data record) but also that the influence of the factors is small. Figure 1 shows the distribution in the number of cigarettes smoked by those with and without full records. The distribution of values is basically the same (clusters around multiples of five up to 20, or, a "pack a day") between the two datasets.

Overall, in spite of the bias from removing heavier smokers with lower apgar scores from the data, the overall number removed is relatively small and the size of the bias (indicated by the coefficients in the linear model) is relatively small.

Table 1: Comparison of data with full records to those with missing data across key variables

full.record	mean.omaps	sd.omaps	mean.fmaps	sd.fmaps	mean.cigar	sd.cigar
FALSE	7.905	1.572	8.880	1.030	3.945	7.422
TRUE	8.117	1.260	9.009	0.707	1.907	5.297

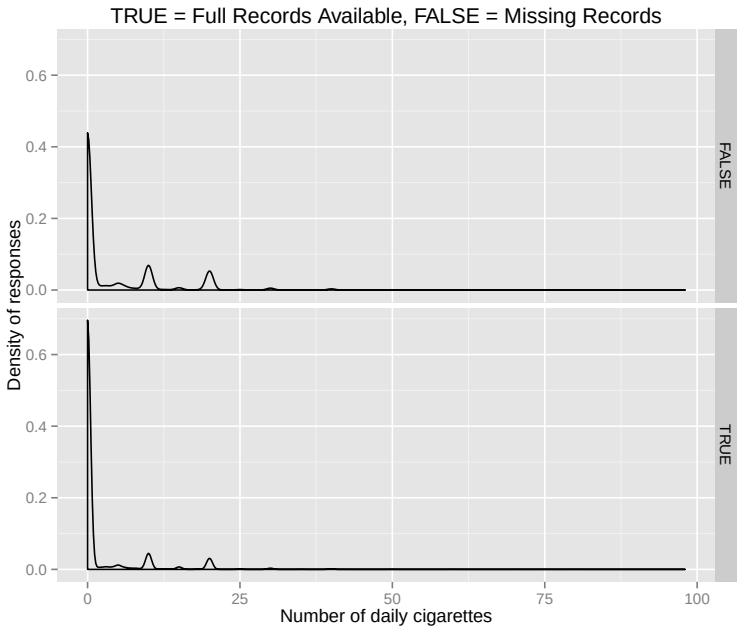

Figure 1: Cigarette use rate by presence of full data record.

Table 2: Linear model results for predicting whether full records are present based on selected variable of interest in the dataset

	Dependent variable:
	full.record
omaps	0.002^{***}
	(0.001)
fmaps	0.007^{***}
	(0.001)
cigar	-0.003^{***}
	(0.0001)
Constant	0.882^{***}
	(0.007)
Observations	119,384
R^2	0.007
Adjusted R^2	0.007
Residual Std. Error	0.195
F Statistic	276.305
Note:	*p<0.1; **p<0.05; ***p<0.01

1.3 Part C

Table 3 is a summary for the remaining data after cleaning.

2 Problem #2

2.1 Part A

The table below shows the mean differences between smoking and non-smoking mothers for one-minute APGAR scores (ompas), five-minute (fmaps), and birth weight in grams (dbrwt). Unconditioned on the other variables, there is no statistically significant difference in APGAR score but a significant difference is present in birth weight[1].

2.2 Part B

The average treatment effect (ATE) of maternal smoking can only be determined by comparing the unadjusted difference in mean birth weight of infants **if their mothers were randomly assigned into treatment (a smoking habit during pregnancy) or the assignment / selection to treatment is as good as random**. This is obviously not possible or even palatable for a variety of practical and ethical reasons to verify with RCT so an alternative approach to identifying the ATE that controls for observables is the next-best option. If we assume that smoking habits are randomly assigned among pregnant mothers, it can be "safe" to use the unadjusted difference in weight as a predictor of ATE without conditioning on observables as long as there are not any significant differences in the smoking and non-smoking groups that also influence birth weight. In the next set of steps we explore other factors that may influence birth weight and if they are also related to smoking status.

ATE using unadjusted differences: If we were to assume that smoking is in fact randomly assigned, the mean difference in birth weight caused by smoking between infants whose mothers smoke and those who do not is 240 grams (with a 95% confidence interval of 230 - 250 grams). Infants whose

[1] Welch Two Sample t-test, alternative hypothesis: true difference in means is not equal to 0; p-value less than 2.2e-16, 95 percent confidence interval: -249.5463 to -231.4093

Table 3: Summary of clean data

variable	var.labels	var	n	mean	sd	se
rectype	record type	1	114610	1.26	0.44	0.00
pldel3	facility of birth recode	2	114610	1.02	0.13	0.00
birattnd	attendant at birth	3	114610	1.20	0.56	0.00
cntocpop	county of occurence	4	114610	1.44	1.14	0.00
stresfip	state of residence	5	114610	41.74	2.17	0.01
dmage	age of mother	6	114610	27.76	5.70	0.02
ormoth	hispanic origin of mother	7	114610	0.09	0.52	0.00
mrace3	race of mother recode	8	114610	1.26	0.66	0.00
dmeduc	detailed educ of mother	9	114610	13.21	2.27	0.01
dmar	marital status of mother	10	114610	1.25	0.43	0.00
adequacy	adequacy of care recode	11	114610	1.30	0.55	0.00
nlbnl	number of live births, now living	12	114610	0.97	1.15	0.00
dlivord	number of live births, now dead	13	114610	1.99	1.17	0.00
dtotord	detail total birth order	14	114610	2.42	1.52	0.00
totord9	total birth order recode	15	114610	2.41	1.46	0.00
monpre	month pregnancy prenatal care began	16	114610	2.50	1.33	0.00
nprevist	total number of prenatal visits	17	114610	11.15	3.52	0.01
dislb	interval since last live birth	18	114610	350.41	362.33	1.07
islb10	interval since last live birth recode	19	114610	3.32	3.19	0.01
dfage	age of father	20	114610	30.06	6.41	0.02
orfath	hispanic origin of father	21	114610	0.09	0.53	0.00
dfeduc	educ of father detail	22	114610	13.28	2.33	0.01
birmon	month of birth	23	114610	6.47	3.39	0.01
weekday	day of week child born	24	114610	4.05	1.88	0.01
dgestat	gestation – detail in weeks	25	114610	39.15	2.44	0.01
csex	sex of child	26	114610	1.49	0.50	0.00
dbrwt	birthweight in grams	27	114610	3373.29	585.17	1.73
dplural	plurality	28	114610	1.03	0.17	0.00
omaps	one minute agpar score	29	114610	8.12	1.26	0.00
fmaps	five minute agpar score	30	114610	9.01	0.71	0.00
clingest	clinical estimate of gestation	31	114610	39.11	2.06	0.01
delmeth5	method of delivery	32	114610	1.55	1.01	0.00
anemia	anemia mother	33	114610	1.99	0.10	0.00
cardiac	cardiac disease mother	34	114610	1.99	0.08	0.00
lung	acute or chronic lung disease mother	35	114610	1.99	0.08	0.00
diabetes	diabetes mother	36	114610	1.97	0.16	0.00
herpes	genital herpes mother	37	114610	1.99	0.08	0.00
chyper	chronic hypertension	38	114610	1.99	0.09	0.00
phyper	pregnancy related hypertension	39	114610	1.97	0.17	0.00
pre4000	previous infant 4000 or more grams	40	114610	1.99	0.12	0.00
preterm	previous preterm infant	41	114610	1.99	0.12	0.00
tobacco	tobacco use during pregnancy	42	114610	1.84	0.37	0.00
cigar	average number of cigarettes per day	43	114610	1.91	5.30	0.02
cigar6	average number of cigarettes per day recode	44	114610	0.35	0.86	0.00
alcohol	alcohol use during pregnancy	45	114610	1.99	0.10	0.00
drink	average number of drinks per week	46	114610	0.03	0.62	0.00
drink5	average number of drinks recode	47	114610	0.02	0.23	0.00
wgain	weight gain	48	114610	30.36	11.88	0.04
full.record*	full record present	49	114610	1.00	0.00	0.00

Table 4: Comparison of key birthing infant health indicators for different maternal smoking status

tobacco	mean.omaps	mean.fmaps	mean.dbrwt
smoker	8.10	9.01	3171
nonsmoker	8.12	9.01	3412
difference	0.017	0.0001	240.5

mother smoked have about 7% lower birth weight than those who did not.

Identifying potential confounding factors: We used deductive logic and graphical exploration to understand factors that may influence birth weight and should be controlled for if the tobacco users / non-users have distributions that are not identical (or very similar) between them. Several (but not all) of the factors that we identified as potential candidates are summarized in the Table **??**. We omitted many that did not show a relationship between the factor and birth weight for brevity. The results show that most of the potential factors related to birth weight do not appear likely to be also related to smoking status.

The factors we identify as having an impact on birth weight AND being related to smoking status are:

- **Maternal Age** is different between the smoking / non-smoking group and is related to birth weight. The median pregnant smoker is two years younger than the median non-smoker. There is also a relationship between age and birth weight (where older mothers up to age 31-32 or so tend to have heavier babies). The relationship between maternal age and birth weight along with the distributions in age for smokers and non-smokers is shown in Figure 2
- **Marital Status** is also different between the smoking and non-smoking groups: single mothers are more likely to smoke in pregnancy. In the whole sample the fraction of women who are married is 75% but in the "smoker" subsample it is only 52%. There is also a relationship between

Table 5: Contingency table for a range of factors by tobacco use status

	smoker	nonsmoker	Combined
	$N = 18266$	$N = 96344$	$N = 114610$
race of mother recode : White	87% (15876)	86% (82748)	86% (98624)
Other	0% (69)	2% (2202)	2% (2271)
Black	13% (2321)	12% (11394)	12% (13715)
sex of child : Male	52% (9462)	51% (49505)	51% (58967)
Female	48% (8804)	49% (46839)	49% (55643)
marital status of mother : Married	52% (9459)	79% (76368)	75% (85827)
Unmarried	48% (8807)	21% (19976)	25% (28783)
plurality : Singleton	98% (17860)	97% (93694)	97% (111554)
Twin	2% (400)	3% (2503)	3% (2903)
Triplet	0% (6)	0% (135)	0% (141)
Quadruplet	0% (0)	0% (12)	0% (12)
alcohol use during pregnancy : Drinker	3% (639)	0% (472)	1% (1111)
Nondrinker	97% (17627)	100% (95872)	99% (113499)
pregnancy related hypertension : 1	2% (369)	3% (3149)	3% (3518)
2	98% (17897)	97% (93195)	97% (111092)
chronic hypertension : 1	1% (120)	1% (764)	1% (884)
2	99% (18146)	99% (95580)	99% (113726)
cardiac disease mother : 1	1% (111)	1% (677)	1% (788)
2	99% (18155)	99% (95667)	99% (113822)
diabetes mother : 1	3% (490)	3% (2587)	3% (3077)
2	97% (17776)	97% (93757)	97% (111533)
previous infant 4000 or more grams : 1	1% (154)	2% (1506)	1% (1660)
2	99% (18112)	98% (94838)	99% (112950)
detailed educ of mother	12 12 12	12 13 16	12 12 16
month pregnancy prenatal care began	2 2 3	2 2 3	2 2 3
age of mother	22 26 30	24 28 32	24 28 32
clinical estimate of gestation	38 40 40	38 40 40	38 40 40
weight gain	20 29 37	24 30 37	23 30 37

$_a\ b\ _c$ represent the lower quartile a, the median b, and the upper quartile c for continuous variables.
Numbers after percents are frequencies.

marital status and birth weight whereby married mothers tend to have slightly heavier babies. These relationships are shown in Figure 3. If one believes that being married leads to less stress for mothers and/or better resources and support it is possible that marital status is a proxy for other determinants of infant weight. However, as is also shown in the Figure (bottom panel) there are different distributions in maternal age between married and unmarried women, with a relationship that suggests age may be a stronger determining factor since single mothers are typically younger than married mothers.

- **Maternal weight gain (less certain)** is related to infant weight at birth but as we note is not as certain in terms of being related strongly with smoking status. The median weight gain is quite similar between the two smoking status groups (29 lbs for smokers vs. 30 lbs for non-smokers) but there is a larger difference in the 25th percentile weight (20 vs. 24 lbs.). The relationship between maternal weight gain and infant weight gain along with the distribution in maternal gain by smoking status is summarized in Figure 4

The position that smoking status is randomly assigned may have some rational basis, but we cannot claim complete randomness. Consider the following line of reasoning:

- This study was conducted in 1993, decades after the link between smoking and poor infant health was established and widely publicized in both the scientific literature and (more importantly) the popular media. While there is a link between maternal educational attainment (smokers tend to have less education, slightly), this can largely be explained by the age of the mothers (many of whom are simply too young to have graduated college, etc.). This education gap could potentially explain a difference in awareness but we posit it is probably a poor proxy. It is reasonable to expect that the vast majority of mothers in the sample know about the link between smoking during pregnancy and poor infant health outcomes, and that the smoking and non-smoking mothers both have the same maternal drive to protect their unborn infants.

- Furthermore, even if the popular exposure were different between smokers and non-smokers, it is standard practice during neonatal care to

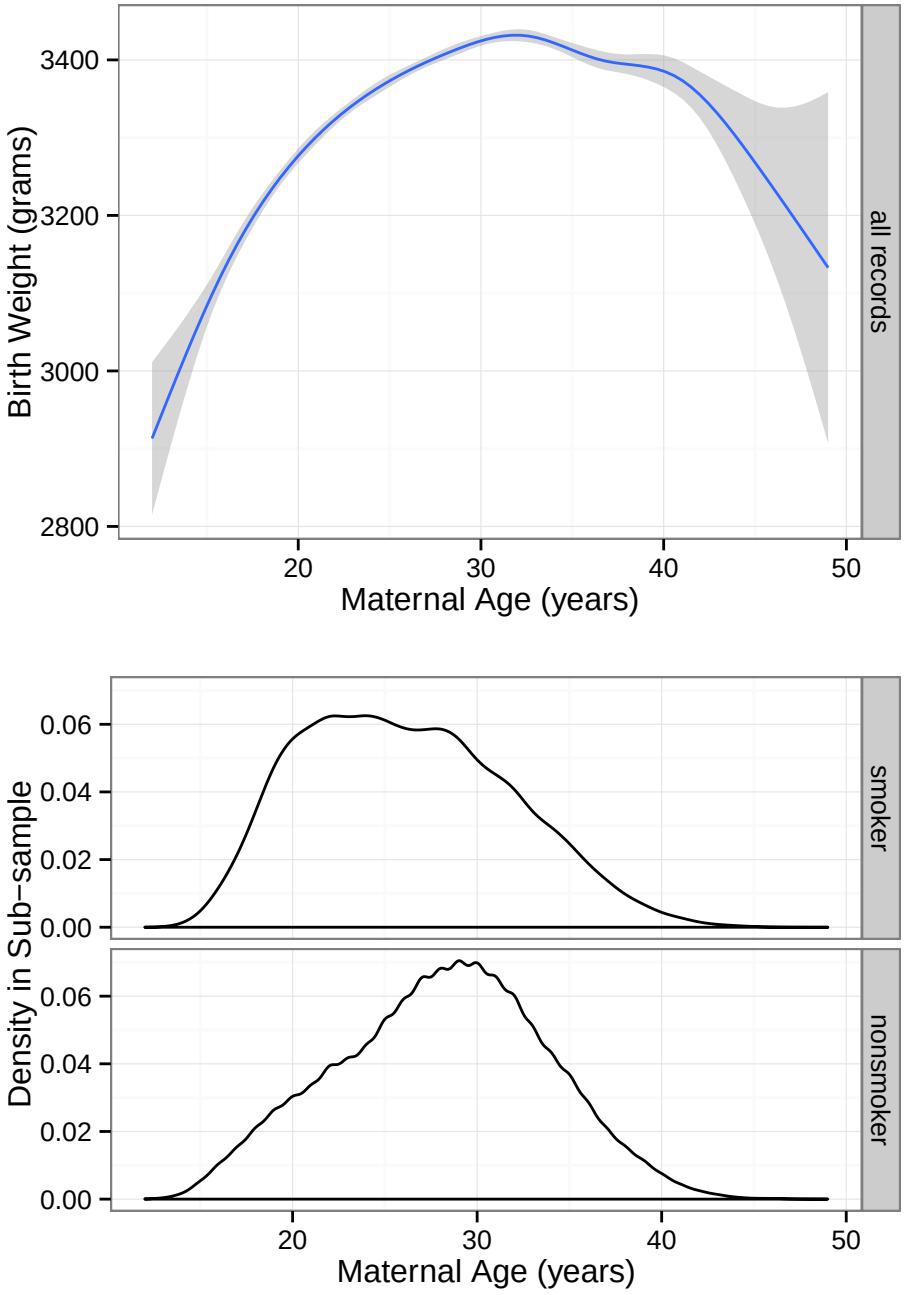

Figure 2: (Top) The relationship between Maternal Age and Birth Weight with a GAM fit to the data and 95% confidence interval estimate in grey. Actual data are omitted to show the average trend more clearly. (Bottom panels) A comparison in the distribution of Maternal Age for smokers and non-smokers shows how smokers tend to be younger mothers.

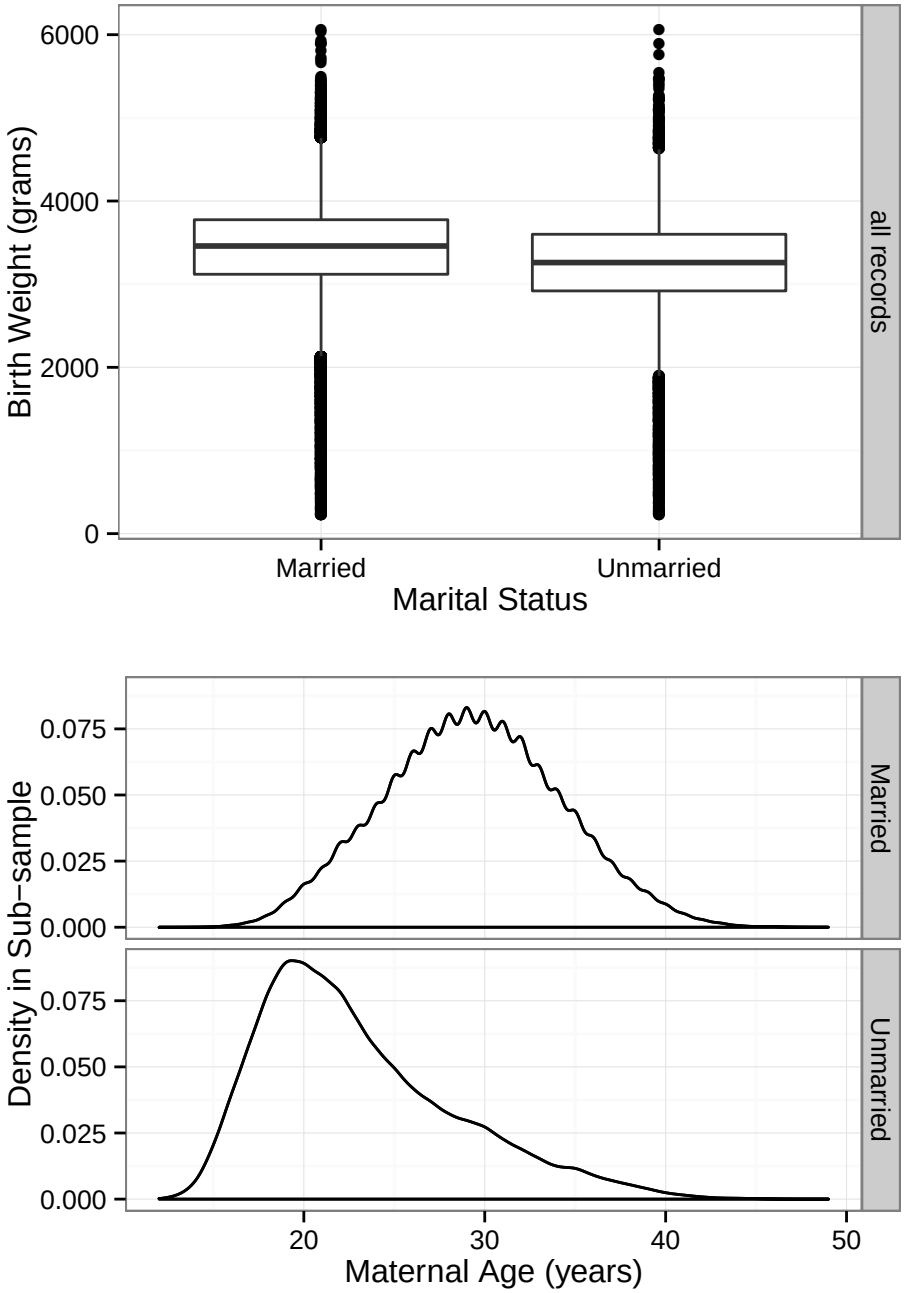

Figure 3: (Top) Boxplots for birth weight by marital status of the mother. (Bottom) Distribution in maternal age between married and unmarried women.

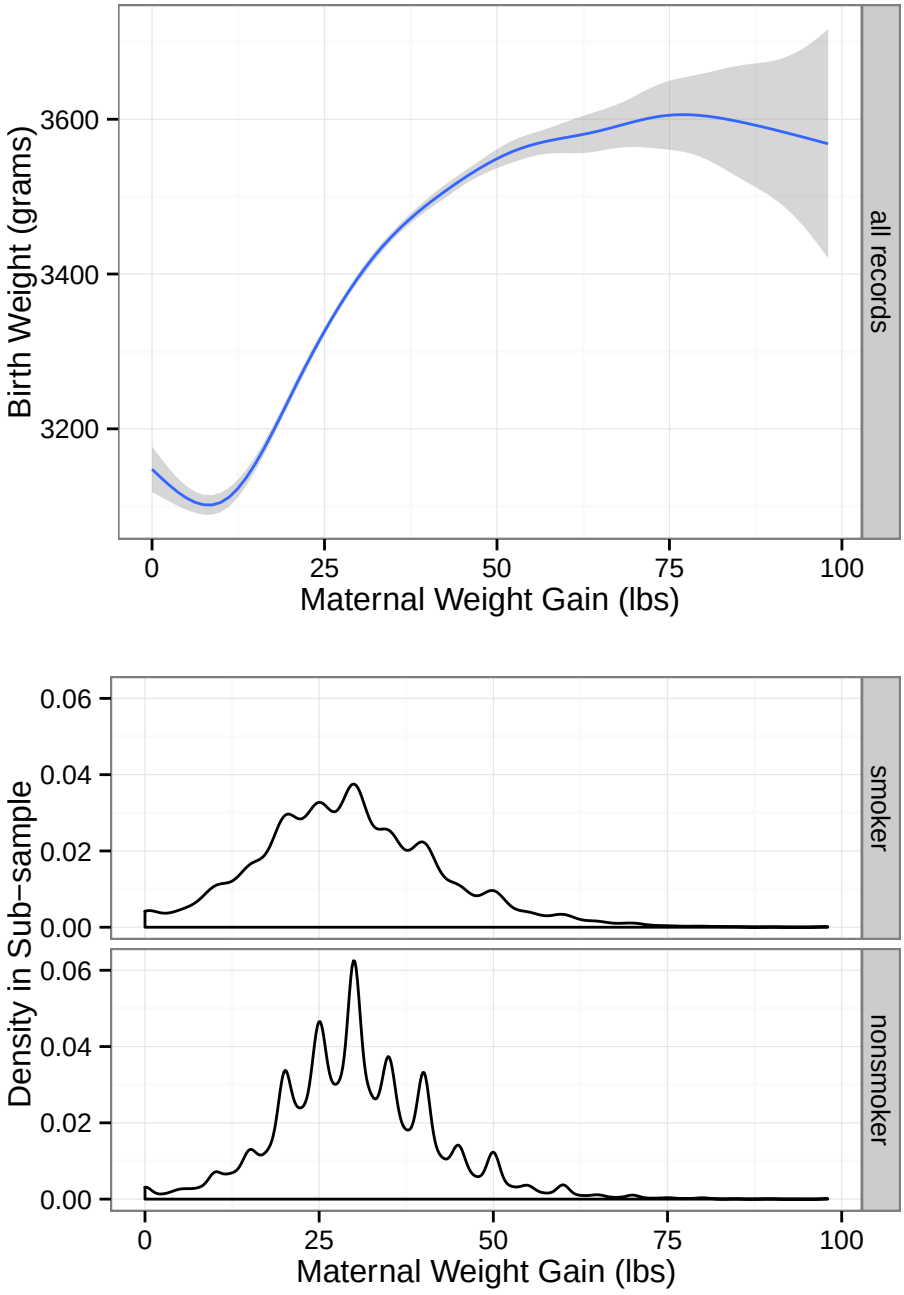

Figure 4: (Top) The relationship between Maternal Weight Gain and Birth Weight with a GAM fit to the data and 95% confidence interval estimate in grey. Actual data are omitted to show the average trend more clearly. (Bottom panels) A comparison in the distribution of Maternal Weight Gain for smokers and non-smokers shows how smokers tend to have (slightly) less weight gain.

receive messages about the value of not smoking. Both smokers and non-smokers presumably received roughly the same level of neonatal care (as measured by the month at which care began).

- If one accepts that awareness about smoking risk and the level of maternal protection drive is the same in both groups, perhaps the only remaining factors are those that underly addiction: genetic predisposition and environmental factors. It is possible, but by no means certain, that the genetic factors (at least) are essentially randomly distributed between people. However, the underlying environmental factors that lead to addiction are not likely randomly distributed and may be correlated with both smoking and other maternal behavior that could lead to lower birth weight.

- Based on the above we posit that there is an element of randomness (genetic factors) associated with smoking but that environmental factors (education, upbringing, social pressures) also contribute strongly to both smoking status and other factors that could cause low birth weight (like pregnancies earlier in life).

Because of the factors we identified it is not defensible to grant the assumption that smoking is randomly assigned in the population. Using unadjusted mean differences is not tenable for obtaining an accurate prediction of ATE.

2.3 Part C

An approach to "correcting" for non-random selection to treatment is identifying predetermined (or exogenous) factors and attempting to control them them. In this case, where we are analyzing data ex post of the "natural experiment" that was documented, we will identify factors that meet three criteria to use as controls: 1) Predetermined or exogenous data, 2) Exhibit some bias in the treatment and control groups, 3) Agnostic to the treatment status, exhibit covariance with birth weight.

Some variables are clearly not predetermined and are endogenous to treatment. In general the predetermined factors are those that can be completely extricated from the treatment, i.e., those that could be changed for a particular individual without changing the treatment category. Factors that are not

predetermined, and therefore cannot be pulled apart from treatment effects, are those that derive at least partly from the smoking status of the mother or underlying factors of smoking status. The ultimate goal is to identify predetermined factors so they can be controlled in the regression to capture only factors that can be effected by the treatment.

It is helpful in this case to define a time at which treatment "begins." Since we might be interested to inform public policy, one could investigate potential benefits of a smoking cessation program for women who are trying to conceive children. Taking this as a benchmark, the experiment we would run, if it were ethical and practical, would be to randomly assign women to either smoke or not smoke just before they became pregnant. Based on this, any outcome that is related to the mother or child's biology after conception is not predetermined (e.g., maternal weight gain). Other factors that could not be predetermined in the experiment are pregnancy-related health concerns like hypertension, the month prenatal care began (since a woman who smokes, and knows it is potentially harmful to the infant, may hesitate to seek medical care), alcohol use during pregnancy (people often smoke and drink socially), gestational age at birth, and other medically related indicators.

There are, however, a range of predetermined factors that could enter the linear model if they meet our other criteria (which are in place to avoid a "crowded / kitchen sink" model that is difficult to describe and interpret): bias in the selection to treatment groups and an obvious influence on birth weight.

Based on our analysis and reasoning, there are two key factors that are predetermined, contribute to birth weight, and also are biased by the "smoking" treatment: Maternal Age and Marital Status. Other factors that are predetermined but do not meet the other two criteria (contributing to weight AND biased by smoking treatment comported to the larger sample) include the sex of the baby, level of prenatal care, age of the father (to the extent that there are genetic causes), gestational time, state of residence, infant plurality, etc. There are also factors that are not predetermined and seem to be (potentially) closely linked with treatment status, such as alcohol use in utero (which was a small fraction of the whole population), maternal weight gain, pregnancy related hypertension, and incidences of lung disease.

2.4 Part D

Selection on observables strategies for teasing out causality guide us to identify all the observable factors (except for the treatment–smoking in this case) that could lead to the outcome (birth weight) and to "correct" for these using statistical techniques before the test for smoking. We identified three additional key factors above that we will use in a set of simple linear models to understand how different factors influence birth weight: Maternal Age, Marital Status, and Maternal Weight Gain. We include maternal weight gain with prejudice on how to interpret it because it too could be linked with smoking status (anecdotal evidence indicates that smoking tends to suppress weight gain and that smoking cessation could lead to temporary weight gain). Table **??** and **??** below summarize the suite of linear models we used to explore the relationships between various potential causal factors and birth weight

The outcome of the models indicates that each of the three additional factors (in addition to tobacco use) is statistically significant in terms of predicting birth weight. However, because maternal weight gain is not a clearly predetermined factor we choose to ignore models that include it (although it is interesting to examine them). The "best" model that only includes Maternal Age and Marital Status as conditioning variables along with tobacco use.

By introducing conditioning variables in a linear regression we downgrade the estimate for average treatment effect from 240 grams to about 200 grams because smokers tended to be younger and unmarried (both significantly decrease birth weight).

Table 6: Birth weight linear models without including Tobacco factors

	Birth Weight		
	(1)	(2)	(3)
Maternal Age	9.536***	2.888***	4.246***
	(0.302)	(0.341)	(0.334)
Marital Status (unmarried)		−183.626***	−175.893***
		(4.479)	(4.384)
Weight Gain			10.042***
			(0.141)
Constant	3,108.613***	3,339.251***	2,994.784***
	(8.558)	(10.190)	(11.081)
N	114,610	114,610	114,610
R^2	0.009	0.023	0.064
Adjusted R^2	0.009	0.023	0.064
Residual Std. Error	582.649 (df = 114608)	578.426 (df = 114607)	566.024 (df = 114606)
F Statistic	996.918*** (df = 1; 114608)	1,346.077*** (df = 2; 114607)	2,629.910*** (df = 3; 114606)

Notes:

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Table 7: Birth weight linear models including Tobacco factors

	Birth Weight		
	(1)	(2)	(3)
Maternal Age	4.058***	2.716***	7.781***
	(0.332)	(0.338)	(0.301)
Marital Status (unmarried)	−141.095***	−146.507***	
	(4.444)	(4.538)	
Weight Gain	9.850***		
	(0.140)		
Tobacco (nonsmoker)	183.678***	195.101***	225.823***
	(4.668)	(4.764)	(4.690)
Constant	2,842.679***	3,170.698***	2,967.483***
	(11.666)	(10.921)	(8.965)
N	114,610	114,610	114,610
R^2	0.077	0.037	0.028
Adjusted R^2	0.077	0.037	0.028
Residual Std. Error	562.240 (df = 114605)	574.242 (df = 114606)	576.845 (df = 114607)
F Statistic	2,386.185*** (df = 4; 114605)	1,469.452*** (df = 3; 114606)	1,667.935*** (df = 2; 114607)

Notes:

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

3 [R] Code

We used R to complete this assignment. The code is below:

```
### This is Peter Alstone & Frank Proulx's solution to ARE213 PS1a
## Data is in the file "ps1.dta"


## working directories --------

# Peter
# setwd("~/Google Drive/ERG/Classes/ARE213/are213/ps1")

# Frank
setwd("~/media/frank/Data1/documents/School/Berkeley/Fall13/ARE213/are213/ps1")

## PACKAGES -------------

library(foreign) #this is to read in Stata data
library(Hmisc)
library(psych)
library(stargazer)
library(ggplot2) # for neato plotting tools
library(plyr) # for nice data tools like ddply
library(car) # "companion for applied regression" - recode fxn, etc.
library(gmodels) #for Crosstabs

# custom functions
source("../util/are213-func.R")
source("../util/watercolor.R") # for watercolor plots


## DATA -------------

ps1.data <- read.dta(file="ps1.dta")
#changed name of object from "data" to avoid ambiguity issues since "data" is often embedded in functions as a general object

print(nrow(ps1.data))



## Problem 1a: Fix missing values --------
## The following are the error codes for each of the 15 variables that need fixing:
# For cardiac - alcohol: "8" means missing record
# 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

# Identify which records have full data, then add a column to indicate full records or not
full.record.flag <- which(ps1.data$cardiac != 9 &
                            ps1.data$cardiac != 8 &
                            ps1.data$lung != 9 &
                            ps1.data$lung != 8 &
                            ps1.data$diabetes !=9 &
                            ps1.data$diabetes !=8 &
                            ps1.data$herpes != 9 &
                            ps1.data$herpes != 8 &
                            ps1.data$chyper != 9 &
                            ps1.data$chyper != 8 &
                            ps1.data$phyper != 9 &
                            ps1.data$phyper != 8 &
                            ps1.data$pre4000 !=9 &
                            ps1.data$pre4000 !=8 &
                            ps1.data$preterm != 9 &
                            ps1.data$preterm != 8 &
                            ps1.data$tobacco != 9 &
                            ps1.data$cigar != 99 &
                            ps1.data$cigar6 !=6 &
                            ps1.data$alcohol != 9 &
                            ps1.data$drink != 99 &
                            ps1.data$drink5 !=5 &
                            ps1.data$wgain !=99
                          )

# Column with flags for full records
ps1.data$full.record <- FALSE # initialize column as F
ps1.data$full.record[full.record.flag] <- TRUE #reassign level to T for full records


# Problem 1b: Describe dropped levels --------

# replace error rows in cigar with NA so they don't interfere with other calcs on influence of dropped values.
error.cigar <- which(ps1.data$cigar == 99)
ps1.data$cigar[error.cigar] <- NA

# compare records on things that (might) matter for this analysis...apgar, smoking, etc.
ps1.compare.records <- ddply(ps1.data, .(full.record), summarize,
                             mean.omaps = mean(omaps),
                             sd.omaps = sd(omaps),
                             mean.fmaps = mean(fmaps),
                             sd.fmaps = sd(fmaps),
                             mean.cigar = mean(cigar, na.rm = TRUE),
                             sd.cigar = sd(cigar, na.rm = TRUE)
                             )

```

```
#--> RESULT: There appears to be a variation in the mean cigarette use
    between groups, but with large standard deviation.

# Print result table for comparison
stargazer(ps1.compare.records, summary=FALSE)

# Plot to explore if missing value people smoke more cigarettes
pdf(file="img/cigar-by-record-type.pdf", width = 7, height = 6)
ggplot(ps1.data, aes(cigar)) + geom_density() + facet_grid(full.record~.) +
    xlab("Number of daily cigarettes") + ylab("Density of responses") +
    ggtitle("TRUE = Full Records Available, FALSE = Missing Records")
dev.off()

# Linear model to see if you can predict whether the data have a full record
     based on cigar, omaps, fmaps
unclean.cig <- lm(full.record ~ cigar, ps1.data)
unclean.cig.om <- lm(full.record ~ omaps + cigar, ps1.data)
unclean.cig.om.fm <- lm(full.record ~ omaps + fmaps + cigar, ps1.data)

# The models seem to indicate you can predict whether there is a full record
     based on apgar and cigarette use....unfortunate.
stargazer(unclean.cig.om.fm)

ps1.data.clean <- subset (ps1.data, full.record == TRUE)
ps1.data.missingvalues <- subset(ps1.data, full.record == FALSE)

print(nrow(ps1.data.clean)) #number of records remaining after cleaning

# Problem 1c: Summary table of clean data, write a new csv-------
var.labels <- attr(ps1.data, "var.labels")
var.labels[length(var.labels)+1] <- "full record present"

ps1.names <- data.frame("labels" = as.data.frame(var.labels))
colnames(ps1.names)[1] <- "labels"

summarytable<-print(describe(ps1.data.clean, skew=FALSE, ranges=FALSE))

latex(title="variable", file="clean-summary.tex" , cbind(var.labels,
    summarytable), caption="Summary of clean Data", vbar=TRUE, size="
    footnotesize")

# stargazer(ps1.data.clean) # Doesn't work as well as the Hmisc version for
    this long table.

write.csv(ps1.data.clean, file = "ps1dataclean.csv")


#Problem 2a Simple difference in APGAR and birth weight -------

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

#change tobacco to factor and label values
ps1.data.clean$tobacco <- as.factor(ps1.data.clean$tobacco)
ps1.data.clean$tobacco <- revalue(ps1.data.clean$tobacco, c( "1" = "smoker",
     "2" = "nonsmoker" ))
```

```
152
153 smoke.impact <- ddply(ps1.data.clean, .(tobacco), summarize,
154                     mean.omaps = mean(omaps),
155                     mean.fmaps = mean(fmaps),
156                     mean.dbrwt = mean(dbrwt)
157                     )
158
159 # conversion to character class for tobacco
160 smoke.impact$tobacco <- as.character(smoke.impact$tobacco)
161 # Add difference row
162 smoke.impact <- rbind(smoke.impact, c("difference", apply(smoke.impact
        [,2:4], 2, diff)))
163
164 stargazer(smoke.impact, summary=FALSE, digits = 2)
165
166 # # alt version 2a-------
167 #
168 # smokers <- subset(ps1.data.clean, tobacco==1)
169 # nonsmokers <- subset(ps1.data.clean, tobacco==2)
170 #
171 # smokerstats <- c(mean(smokers$omaps), mean(smokers$fmaps), mean(smokers$
        dbrwt))
172 # nonsmokerstats <- c(mean(nonsmokers$omaps), mean(nonsmokers$fmaps), mean(
        nonsmokers$dbrwt))
173 # meandif <- nonsmokerstats - smokerstats
174 #
175 # smoketable <- matrix(c(smokerstats, nonsmokerstats, meandif), ncol=3,
        byrow=FALSE)
176 # colnames(smoketable) <- c("Mean Value (Infants with Smoker Mothers)", "
        Mean Value (Infants with Non-Smoker Mothers)", "Mean Difference between
        control and treatment")
177 # rownames(smoketable) <- c("one minute APGAR score", "five munute APGAR
        score", "birthweight")
178 # smoketable <- as.data.frame(smoketable)
179 #
180 #
181 # stargazer(smoketable, title = "Mean values of health figures in Infants
        with Smoker and Non-Smoker Mothers", type="latex")
182
183 # Problem 2b -------
184
185 #recode variables
186
187 ps1.data.clean$mrace3 <- as.factor(ps1.data.clean$mrace3)
188 ps1.data.clean$mrace3 <- revalue(ps1.data.clean$mrace3, c( "1" = "White", "2
        " = "Other", "3" = "Black" ))
189
190 ps1.data.clean$csex <- as.factor(ps1.data.clean$csex)
191 ps1.data.clean$csex <- revalue(ps1.data.clean$csex, c( "1" = "Male", "2" = "
        Female"))
192
193 ps1.data.clean$dplural <- as.factor(ps1.data.clean$dplural)
194 ps1.data.clean$dplural <- revalue(ps1.data.clean$dplural, c("1" = "Singleton
        ", "2" = "Twin", "3"= "Triplet", "4" = "Quadruplet", "5" = "Quintuplet+"
        ))
195
196 ps1.data.clean$alcohol <- as.factor(ps1.data.clean$alcohol)
```

```
ps1.data.clean$alcohol <- revalue(ps1.data.clean$alcohol, c("1" = "Drinker",
     "2" = "Nondrinker", "9" = "Unk."))

ps1.data.clean$dmar <- as.factor(ps1.data.clean$dmar)
ps1.data.clean$dmar <- revalue(ps1.data.clean$dmar, c("1" = "Married", "2" =
     "Unmarried"))

# T-tests for relationships.

print( t.test( omaps ~ tobacco, data = ps1.data.clean))
print( t.test( fmaps ~ tobacco, data = ps1.data.clean))
print( t.test( dbrwt ~ tobacco, data = ps1.data.clean))

#visual representation of relationship:

ps1.data.clean$all <- "all records" # flag for facets with all the same.

# birth weight - age
pdf(file="img/bw-age.pdf", width=5, height=7)
bw.age <- ggplot(ps1.data.clean, aes(dmage, dbrwt))
bw.age <- bw.age +
  stat_smooth() +
  theme_bw() +
  xlab("Maternal Age (years)") +
  ylab("Birth Weight (grams)") +
  facet_grid(all~.)

split.age <- ggplot(ps1.data.clean, aes(x=dmage))
split.age <- split.age +
  geom_density() +
  theme_bw() +
  xlab("Maternal Age (years)") +
  ylab("Density in Sub-sample") +
  facet_grid(tobacco~.)

arrange_ggplot2(bw.age, split.age, ncol=1)
dev.off()


# birth weight - marriage
pdf(file="img/bw-mar.pdf", width=5, height=7)

bw.mar <- ggplot(ps1.data.clean, aes(factor(dmar),dbrwt))
bw.mar <- bw.mar +
  geom_boxplot() +
  theme_bw() +
  xlab("Marital Status") +
  ylab("Birth Weight (grams)") +
  facet_grid(all~.)

split.mar <- ggplot(ps1.data.clean, aes(x=dmage))
split.mar <- split.age +
  geom_density() +
  theme_bw() +
  xlab("Maternal Age (years)") +
  ylab("Density in Sub-sample") +
  facet_grid(dmar~.)
```

```
252
253 arrange_ggplot2(bw.mar, split.mar, ncol=1)
254
255 dev.off()
256
257
258 # birth weight - maternal weight gain
259 pdf(file="img/bw-gain.pdf", width=5, height=7)
260 bw.gain <- ggplot(ps1.data.clean, aes(wgain, dbrwt))
261 bw.gain <- bw.gain +
262   stat_smooth() +
263   theme_bw() +
264   xlab("Maternal Weight Gain (lbs)") +
265   ylab("Birth Weight (grams)") +
266   facet_grid(all~.)
267
268 split.gain <- ggplot(ps1.data.clean, aes(x=wgain))
269 split.gain <- split.gain +
270   geom_density() +
271   theme_bw() +
272   xlab("Maternal Weight Gain (lbs)") +
273   ylab("Density in Sub-sample") +
274   facet_grid(tobacco~.)
275
276 arrange_ggplot2(bw.gain, split.gain, ncol=1)
277 dev.off()
278
279 ## CrossTabs
280 ps1.xtab.data <- ps1.data.clean
281 # Add labels from ps1.data
282 for(i in 1:length(names(ps1.data))){
283   label(ps1.xtab.data[[i]]) <- attr(ps1.data, "var.labels")[i]
284 }
285
286 # not-that-useful function for generating a list of crosstabs...
287 xtab.create <- function(data, const.col, cross.col, counts = FALSE){
288 # returns a data frame with percentiles of each cross column holding const
      column consant
289 #
290 # data = a data frame
291 # const.col = column to be held constant (usually treatment)
292 # cross col = columns to vary (a concatenated list of character strings or
      indicies
293
294 # ERRORS
295 # TODO: Error handling
296   # race of mother
297 xtab.out <- list()
298
299 for(factor in cross.col){
300   xtab.factor <- CrossTable(data[[factor]],data[[const.col]])
301   store.ver <- reshape(data=as.data.frame(xtab.factor$prop.row),idvar="x",
        direction="wide",timevar="y")
302   if(counts){
303
304
305   }
```

```
306   xtab.out[[factor]] <- store.ver
307 }
308 return(xtab.out)
309 }
310
311
312
313 # # Improved labels for table (optional)
314 # label(ps1.data.clean$dmage) <- "Maternal Age (yr)"
315 # label(ps1.data.clean$tobacco) <- "Tobacco Use Status"
316 # label(ps1.data.clean$mrace3) <- "Maternal Race"
317 # label(ps1.data.clean$csex) <- "Infant Sex"
318 # label(ps1.data.clean$dplural) <- "Infant Plurality"
319 # label(ps1.data.clean$clingest) <- "Gestational Age (weeks)"
320 # label(ps1.data.clean$alcohol) <- "Alcohol Use Status"
321 # label(ps1.data.clean$phyper) <- "Preg. Hypertension"
322
323 # Grouped crosstabs using Hmisc
324
325 latex(summary( tobacco  ~
326                  mrace3 +
327                  csex +
328                  dmar +
329                  dplural +
330                  alcohol +
331                  phyper +
332                  chyper +
333                  cardiac +
334                  diabetes +
335                  pre4000 +
336                  dmeduc +
337                  monpre +
338                  dmage +
339                  clingest +
340                  wgain,
341                data=ps1.xtab.data,
342                method="reverse",
343                overall=TRUE, long=TRUE
344                ),
345       title = "crosstab-tobacco",
346       label = "tab:xtabTobacco",
347       caption = "Contingency table for a range of factors by tobacco use
              status",
348       exclude1=F
349       )
350
351 # Problem 2c --------
352 # This one is all in latex doc.  Just describing things.
353
354 # Problem 2d ---------
355
356 sm.age <- lm(dbrwt ~ dmage, ps1.data.clean)
357
358 sm.age.mar <- lm(dbrwt ~ dmage + dmar, ps1.data.clean)
359
360 sm.a.m.w <- lm(dbrwt ~ dmage + dmar + wgain, ps1.data.clean)
361
```

```
sm.a.m.w.t <- lm(dbrwt ~ dmage + dmar + wgain + tobacco, ps1.data.clean)

sm.a.m.t <- lm(dbrwt ~ dmage + dmar + tobacco, ps1.data.clean)

sm.a.t <- lm(dbrwt ~ dmage + tobacco, ps1.data.clean)

sm.axm.t <- lm(dbrwt ~ dmage * dmar + tobacco, ps1.data.clean) #not used -
    look for cross of age:mar

# table without tobacco
stargazer(sm.age, sm.age.mar, sm.a.m.w,
          type="latex",
          covariate.labels = c("Maternal Age", "Marital Status (unmarried)",
               "Weight Gain"),
          align = TRUE,
          style="qje",
          single.row = FALSE,
          font.size="footnotesize",
          dep.var.labels = "Birth Weight",
          out = "combinedReg-noTob.tex"
          )

# table with tobacco
stargazer(sm.a.m.w.t, sm.a.m.t, sm.a.t,
          type="latex",
          covariate.labels = c("Maternal Age", "Marital Status (unmarried)",
               "Weight Gain", "Tobacco (nonsmoker)"),
          align = TRUE,
          style="qje",
          single.row = FALSE,
          font.size="footnotesize",
          dep.var.labels = "Birth Weight",
          out = "combinedReg-withTob.tex"
)
```

ps1a1.R

```
# Econometrics helper functions for [R]
#
# Peter Alstone and Frank Proulx
# 2013
# version 1
# contact: peter.alstone AT gmail.com

# Category: Data Management -------------


# Category: Data Analysis ----------------

# Function: Find adjusted R^2 for subset of data
# This requires a completed linear model...pull out the relevant y-values
    and residuals and feed them to function
# [TODO @Peter] Improve function so it can simply evaluate lm or glm object,
     add error handling, general clean up.
adjr2 <- function(y,resid){
  r2 <- 1-sum(resid^2) / sum((y-mean(y))^2)
  return(r2)
```

```
} #end adjr2


# Category: Plots and Graphics -------------

## Function for arranging ggplots. use png(); arrange(p1, p2, ncol=1); dev.
    off() to save.
require(grid)
vp.layout <- function(x, y) viewport(layout.pos.row=x, layout.pos.col=y)
arrange_ggplot2 <- function(..., nrow=NULL, ncol=NULL, as.table=FALSE) {
  dots <- list(...)
  n <- length(dots)
  if(is.null(nrow) & is.null(ncol)) { nrow = floor(n/2) ; ncol = ceiling(n/
      nrow)}
  if(is.null(nrow)) { nrow = ceiling(n/ncol)}
  if(is.null(ncol)) { ncol = ceiling(n/nrow)}
  ## NOTE see n2mfrow in grDevices for possible alternative
  grid.newpage()
  pushViewport(viewport(layout=grid.layout(nrow,ncol) ) )
  ii.p <- 1
  for(ii.row in seq(1, nrow)){
    ii.table.row <- ii.row
    if(as.table) {ii.table.row <- nrow - ii.table.row + 1}
    for(ii.col in seq(1, ncol)){
      ii.table <- ii.p
      if(ii.p > n) break
      print(dots[[ii.table]], vp=vp.layout(ii.table.row, ii.col))
      ii.p <- ii.p + 1
    }
  }
}
```

../util/are213–func.R